QUARTERLY OF APPLIED MATHEMATICS

Vol. XVII OCTOBER, 1959 No. 3

RESTRICTIONS IMPOSED UPON THE UNIT STEP RESPONSE OF LINEAR PHASE SHIFT NETWORKS*

BY

PAUL M. CHIRLIAN

New York University

Summary. Bounds have been placed upon the 10–90% rise time and overshoot, of the unit step response of phase distortion free networks. Two bounds on the rise time are given, one, in terms of the area under the amplitude function curve, and the other in terms of the area under the square of this curve. Two bounds on the overshoot are given in terms of these areas and the phase slope or its square root.

Best possible bounds on the rise time and overshoot are obtained when the amplitude function is itself bounded and approaches zero at least as fast as ω^{-n}, $n > 0$ in the high frequency region. These bounds may be evaluated readily from tabulated data.

Introduction. The transient response of a network is often substantially improved by the introduction of phase equalization. However, it should be determined whether such phase correction will actually provide sufficient improvement to warrant its use. Computations of the transient response are often tedious. If certain key quantities such as the 10–90% rise time and the overshoot, of the unit step response, of the phase corrected network could be readily estimated, then the advisability of phase correction could be determined. Such readily evaluated bounds are presented here.

Since phase distortion free networks have a linear phase characteristic, it will be assumed throughout this paper that the transfer function of the network in question is of the form $T(\omega)e^{-jk\omega}$ where $T(\omega) \geq 0$ and k is the constant phase slope. Low pass structures will be considered and for convenience $T(\omega)$ will be normalized so that $T(0) = 1$. The unit step response will be written as $A(t)$.

The 10–90% rise time of the unit step responses. Bounds on the 10–90% rise time can be obtained by manipulation of the expression for the unit step response

$$A(t) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{T(\omega)}{\omega} \sin \omega(t - k)\, d\omega. \tag{1}$$

Utilizing the fact that the unit step response of a linear phase network possesses odd symmetry about the 50% point, which occurs at $t = k$, the 10–90% rise time r_t can be obtained by solution of the equation

*Received June 2, 1958; revised manuscript received October 27, 1958. This paper is based upon a portion of a thesis which has been accepted by the faculty of the Graduate Division, College of Engineering, New York University, in partial fulfillment of the requirements for the degree of Doctor of Engineering Science.

$$\frac{4}{5} = \frac{2r_t}{\pi} \int_0^\infty \frac{T(\omega) \sin [\omega r_t/2]}{\omega r_t} \, d\omega. \tag{2}$$

It is assumed here that the 10 and 90% points fall on opposite sides of the 50% point. If this is not the case, then the response will fall below 10% after $t = k$. Such a highly oscillatory response is not suitable for most practical applications and will not be considered here.

Equation (2) is difficult to evaluate. In fact, the rise time can, in general be more readily evaluated from a plot of the actual transient response. However, Eq. (2) can be used to obtain some readily evaluated bounds.

Since

$$\left| \frac{\sin x}{x} \right| \leq 1,$$

$$r_t \geq \frac{4\pi}{5} \left[\int_0^\infty T(\omega) \, d\omega \right]^{-1}. \tag{3}$$

Often $\int_0^\infty T(\omega) d\omega$ does not exist, which renders inequality (3) useless. Even if the integral does exist, the lower bound may be unduly small.

A somewhat different approach may be used to obtain a bound that exists in all physical cases where parasitic capacitance limits the high frequency response.

Apply Schwartz' inequality to Eq. (2).

$$\frac{4}{5} \leq \frac{2}{\pi} r_t \left\{ \int_0^\infty [T(\omega)]^2 \, d\omega \int_0^\infty \left[\frac{\sin (\omega r_t/2)}{\omega r_t} \right]^2 d\omega \right\}^{\frac{1}{2}}$$

but,

$$\int_0^\infty \left[\frac{\sin (\omega r_t/2)}{\omega r_t} \right]^2 d\omega = \frac{\pi}{4r_t}.$$

Thus,

$$r_t \geq \frac{16}{25} \pi \left\{ \int_0^\infty [T(\omega)]^2 \, d\omega \right\}^{-1}. \tag{4}$$

In many instances the bound presented by relation (4) will be much stronger than that stated in relation (3). Of course, there are also conditions where the converse is true.

If some relatively weak conditions, which are obtained in practical situations, are imposed upon $T(\omega)$, then a very readily evaluated greatest lower bound on the rise time may be obtained.

Consider that $T(\omega) \leq \epsilon(\omega_c/\omega)^n$, $n > 0$ for $\omega > \omega_c$. That is $T(\omega_c) \leq \epsilon$ and $T(\omega)$ falls off as ω^{-n}, for $\omega > \omega_c$. (For further discussion of such approximations see [1]).

Then the unit step response can be written in the form

$$A(t) = \frac{1}{2} + \frac{1}{\pi} \int_0^{\omega_c} \frac{T(\omega) \sin \omega(t - k)}{\omega} \, d\omega + \delta, \tag{5}$$

where

$$| \delta | \leq \frac{\epsilon}{n\pi}. \tag{6}$$

Thus, if $|\delta|$ is sufficiently small, the transient response can be assumed to be unaffected by frequency components greater than ω_c.

If the transfer function is such that frequency components greater than ω_c can be neglected, and, in addition, if $T(\omega)$ is bounded so that $T(\omega) \leq M$, then the following procedure leads to a readily evaluated bound.

From Eq. (2)

$$\frac{4}{5} = \frac{2}{\pi} \int_0^{\omega_c r_t/2} T(2x/r_t) \frac{\sin x}{x}\, dx.$$

If $\omega_c r_t/2 \leq \pi$, then the following bound is obtained

$$\frac{4}{5} \leq \frac{2}{\pi} M Si(\omega_c r_t/2)$$

or

$$r_t \geq \frac{2}{\omega_c} \text{Inv. } Si[4\pi/(10M)]. \tag{7}$$

Where Inv. $Si[x]$ is defined by the following relations if $x = \int_0^y [\sin z/z]\, dz = Si(y)$ then $y = \text{Inv. } Si(x)$.

Note that Inv. $Si[4\pi/(10M)]$ is single valued since $M \geq 1$ and $[4\pi/(10M)] \leq Si[2\pi]$ which is the least minimum of $Si[x]$.

In the derivation of relation (7) it was assumed that $\omega_c r_t/2 \leq \pi$. If for some $T(\omega)$, $\omega_c r_t/2 > \pi$, then relation (7) will still be valid, since the result is given in the form of a lower bound and, thus, larger values of r_t are allowed.

It can be shown that values of r_t given by the equality sign of relation (7) exist, and hence, this is the best possible bound. To do this, use the transfer function

$$T(\omega) = \lim_{\epsilon \to 0} \begin{cases} 1, \omega < \epsilon \\ M, \epsilon \leq \omega \leq \omega_c \\ 0, \omega > \omega_c \end{cases}.$$

Although such a transfer function cannot be realized, it can be approximated arbitrarily closely.

Note that an increase in the value of M will reduce the bound on the rise time. However, since $T(0) = 1 \leq M$, large values of M can result in large overshoots.

The overshoot of the unit step response. The overshoot of the unit step response will be defined in the usual way

$$o_s = \text{max. } A(t) - 1.$$

If the maximum value of $A(t)$ occurs at time t_m then manipulation of Eq. (1) yields

$$o_s = -\frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} T\left(\frac{x}{t_m - k}\right) \frac{\sin x}{x}\, dx. \tag{8}$$

In general, this equation cannot be used to obtain an overshoot since t_m the time of its occurrence is not known. However, a bound can be obtained. Since $|\sin x/x| \leq 1$, Eq. (8) leads to

$$o_s \leq -\frac{1}{2} + \frac{t_m - k}{\pi} \int_0^\infty T(\omega)\, d\omega.$$

Of course, this equation still contains the unknown time t_m . The bound attains its maximum value when $t_m - k$ is a maximum. Since the response of a linear phase network is symmetrical about 50% point, and in a physically realizable case is zero for $t < 0$, the maximum value of $t_m - k$ is k. Thus,

$$o_s \leq -\frac{1}{2} + \frac{k}{\pi} \int_0^\infty T(\omega)\, d\omega. \tag{9}$$

The Schwartz' inequality provides a second bound which may yield better results than the above if $\int_0^\infty T(\omega)d\omega$ is unduly large, or does not converge.

This is

$$o_s \leq -\frac{1}{2} + \left\{ \frac{k}{2\pi} \int_0^\infty [T(\omega)]^2\, d\omega \right\}^{1/2}. \tag{10}$$

If $T(\omega)$ is such that $| \delta |$ given by relation (6) is negligible, so that the components of $T(\omega)$ for $\omega > \omega_c$ can be neglected and, in addition, $T(\omega)$ is bounded so that $T(\omega) \leq M$, then a bound on the overshoot which requires no integration can be obtained.

Equation (8) leads to

$$o_s = -\frac{1}{2} + \frac{1}{\pi} \int_0^{\omega_c(t_m - k)} T\left(\frac{x}{t_m - k}\right) [\sin x/x]\, dx.$$

To bound this expression, consider a comb like $T(\omega)$ such that $T[x/(t_m - k)] = 0$ if $\sin x < 0$ and $T[x/(t_m - k)] = M$ if $\sin x \geq 0$. The largest value of 0_s then occurs at $t_m - k\,|_{\max} = k$. Thus,

$$o_s \leq -\frac{1}{2} + \frac{M}{\pi} \{ Si(\pi) + [Si(3\pi) - Si(2\pi)] + \cdots \}.$$

The final term in this series depends upon the numerical value of $\omega_c k$ so that the complete form of the bound may be stated in the following way.

If $n\pi \leq \omega_c k < (n+1)\pi$, $n = 0, 1, 2, 3, \cdots$ and $T(\omega) \leq M$, the following bound on the overshoot is obtained.

$$o_s \leq -\frac{1}{2} + \frac{M}{\pi} \sum_{g=1}^{n} (-1)^{g+1} Si(g\pi), \qquad n = 1, 3, 5, \cdots$$

$$o_s \leq -\frac{1}{2} + \frac{M}{\pi} \left\{ Si(\omega_c k) + \sum_{g=1}^{n} (-1)^{g+1} Si(g\pi) \right\}, \qquad n = 2, 4, \cdots \tag{11}$$

$$o_s \leq -\frac{1}{2} + \frac{M}{\pi} Si(\omega_c k), \qquad n = 0.$$

It can be shown that these are the best possible bounds by using the comb like transfer function discussed previously.

In order to aid in the computation of these bounds and to provide a means of rapidly estimating the overshoot the following table is included.

TABLE I

$$\sum_{g=1}^{n} (-1)^{g+1} Si(g\pi) \text{ for odd } n$$

n	$\sum_{g=1}^{n} (-1)^{g+1} Si(g\pi)$
1	1.8519
3	2.1085
5	2.2504
7	2.3484
9	2.4234
11	2.4840
13	2.5350
15	2.5789

It should be noted that if the term $| \delta |$ defined in relation (6) is not negligible, the bounds given by relation (10) can still be used if they are increased by $| \delta |$.

Conclusion. Two bounds have been placed upon the 10–90% rise time of phase distortion free networks. These are

$$r_t \geq \frac{4}{5}\pi\left[\int_0^\infty T(\omega)\, d\omega\right]^{-1}$$

and

$$r_t \geq \frac{16}{25}\pi\left\{\int_0^\infty [T(\omega)]^2\, d\omega\right\}^{-1}.$$

The following bounds have been placed upon the overshoot.

$$o_s \leq -\frac{1}{2} + \frac{k}{\pi}\int_0^\infty T(\omega)\, d\omega,$$

$$o_s \leq -\frac{1}{2} + \left\{\frac{k}{2\pi}\int_0^\infty [T(\omega)]^2\, d\omega\right\}^{1/2}.$$

The shape of the transfer characteristic will determine which of these bounds is the best and should be used.

If $T(\omega)$ falls off so that its frequency components can be neglected for $\omega > \omega_c$ and in addition, if $T(\omega) \leq M$, then bounds may be obtained which require no integration. These bounds are the best possible and are for the rise time

$$r_t \geq \frac{2}{\omega_c} \text{Inv. } Si\left[\frac{4\pi}{10M}\right]$$

and for the overshoot

$$o_s \leq -\frac{1}{2} + \frac{M}{\pi}\sum_{g=1}^{n} (-1)^{g+1} Si(g\pi), \qquad n = 1, 3, 5, \cdots$$

$$o_s \leq -\frac{1}{2} + \frac{M}{\pi}\left\{Si(\omega_c k) + \sum_{g=1}^{n} (-1)^{g+1} Si(g\pi)\right\}, \qquad n = 2, 4, 6, \cdots$$

$$o_s \leq -\frac{1}{2} + \frac{M}{\pi} Si(\omega_c k), \qquad n = 0,$$

where the constant n is determined from the relation

$$n\pi \leq \omega_c k < (n+1)\pi, \qquad n = 0, 1, 2, 3, \cdots .$$

Acknowledgement. The author is indebted to Professor James H. Mulligan Jr. for the many valuable suggestions and criticisms which were made during the course of this work.

The author also wishes to express his appreciation to Professor Armen H. Zemanian and Professor Charles F. Rehberg for their many valuable comments.

Appendix A. Application of the work of Zemanian to linear phase networks. Zemanian [2] has imposed certain restrictions on the unit step response of a physically realizable one port network in terms of the real part of the impedance function $R(\omega)$. These conditions, when modified slightly, apply to the amplitude function of linear phase networks.

Comparison of the expression for the unit step response of a network

$$A(t) = \frac{2}{\pi}\int_0^\infty \frac{R(\omega)}{\omega}\sin \omega t \, d\omega$$

with Eq. (1) indicates the similarity of the roles played by $R(\omega)$ and $T(\omega)$. Some examples of these dual roles will now be given. It has been shown by Zemanian that:

(1) If the unit step response is to be a monotonically increasing function of time, it is necessary that

$$R(0) \geq R(\omega) \quad \text{for all} \quad \omega.$$

(2) If $R(\omega)$ is a monotonically decreasing function of ω which approaches zero as ω approaches infinity, then the overshoot is bounded by

$$o_s \leq -1 + \frac{2}{\pi}Si(\pi) = 0.1790.$$

These may be written for the case of linear phase networks as

(1) If the unit step response is to be a monotonically increasing function of time, it is necessary that $T(0) \geq T(\omega)$ for all ω.

(2) If $T(\omega)$ is a monotonically decreasing function of ω which approaches zero as ω approaches infinity, then the overshoot is bounded by

$$o_s \leq -\frac{1}{2} + \frac{1}{\pi}Si(\pi) = 0.0895.$$

In a similar way other restrictions imposed upon the real part of an impedance function can be carried over to the amplitude function of linear phase networks.

References

1. P. M. Chirlian, *An investigation of linear phase networks*, New York University, Eng. Sc. D. thesis, Department of Electrical Engineering, May, 1956. Also, P. M. Chirlian, *Bounds on the error in the unit step response of a network*, Quart. Appl. Math. **16**, 432-435 (1959)
2. A. H. Zemanian, *Bounds existing on time and frequency responses of various types of networks*, Proc. IRE **42**, 835-839 (May 1954). Also, A. H. Zemanian, *Further bounds existing on the transient responses of various types of networks*, Proc. IRE **43**, 322-326 (Mar. 1955)

ON THE APPLICATION OF DYNAMIC PROGRAMMING TO A CLASS OF IMPLICIT VARIATIONAL PROBLEMS*

BY

RICHARD BELLMAN (*The Rand Corporation*)

AND

JOHN M. RICHARDSON (*Hughes Aircraft Company*)

1. Introduction. A large and important class of variational problems has the following form. Given a vector equation of the form

$$\frac{dx}{dt} = g(x, y), \qquad x(0) = c, \tag{1}$$

where x is an N-dimensional vector, we wish to determine an m-dimensional vector y so as to minimize a given criterion functional

$$J(y) = \int_0^T h(x, y)\, dt, \tag{2}$$

where $h(x, y)$ is a given scalar function.

The vector y may be subject to constraints of the form

$$r_i(x, y) \leq 0, \qquad i = 1, 2, \cdots, q. \tag{3}$$

In problems involving "terminal control," we meet the problem of minimizing a function only of the final state

$$I(y) = k[x(T)]. \tag{4}$$

A problem of this nature occurs when we wish to have the system in some specified state $x_0(T)$ at time T, without caring how the system gets there. This is usually an idealization, in the sense that a more realistic problem will involve a combination of a criterion of the type appearing in (2) together with some measure of the value of the final state.

As has been shown in some recent publications, see [1], where further references may be found, a variety of problems of this nature arising in economic and engineering control processes may be solved computationally by combining the theory of dynamic programming with modern digital computers.

In recent years, problems of less explicit nature have become more frequent. Thus, for example, what is called the "bang-bang" control problem requires that y be chosen so that the system tend to a specified equilibrium state as rapidly as possible; see [2].

The upper limit of integration is thus not predetermined, but is rather a function of the choice of the vector y. In place of a formulation in precise analytic terms of the type appearing in (2) or (4), we encounter an implicit criterion of the following type:

"When x satisfies a set of conditions C_1, C_2, $\cdots$, C_p for the first time, we want a given scalar function of x to be as small as possible."

*Received June 9, 1958.

A particular example of a problem of this nature, equivalent to one we shall discuss in more detail below, is one in which we require that a preassigned function be a minimum for the first value of T for which $x_1(T) = a_1$, a given value.

A number of quite interesting existence and uniqueness questions arise in conjunction with problem statements of the foregoing kind. These will be discussed at some time in the future. Here we are interested in describing a technique which can be used to obtain computational solutions via the functional equation path of dynamic programming.

The problem becomes of even more interesting nature if we insert some stochastic influences into the process. Let the governing equation be

$$\frac{dx}{dt} = g(x, y, r), \qquad x(0) = c, \tag{5}$$

where r is a random vector. We now wish to minimize an expected deviation, or say the probability that the deviation exceeds a given critical value.

Once again, let us point out that the rigorous groundwork for these questions remains to be laid. However, as we shall see below, we have a simple method for postponing this type of investigation.

As is to be expected, certain simplifications are possible if the underlying equations are linear, i.e. of the form

$$x_{n+1} = Ax_n + y_n + r_n, \qquad x_0 = c, \tag{6}$$

and the criteria quadratic. We shall discuss these cases in some detail since they are of some importance in connection with the application of the method of successive approximations.

Throughout, our aim will be to illustrate the applicability of the functional equation technique of dynamic programming to the computational solution of questions of this kind which appear in many ways to be outside the domain of the classical calculus of variations.

2. Preliminaries. Since, as mentioned above, we are primarily interested in a computational solution of implicit variational problems of the type described in the foregoing section, we shall pose our problem in discrete terms. The recurrence relations we derive will then be ready for use in a digital computer.

In place of the differential relation of (1.4), consider the difference equation

$$x_{n+1} = g(x_n, y_n, r_n), \qquad x_0 = c, \qquad n = 0, 1, \cdots, N. \tag{1}$$

One of the advantages of formulating problems in this fashion is that there are now no conceptual difficulties concerning the meaning of random functions or the existence of minimizing functions. In return, sometime or other we must show that the limit of the discrete process exists, and, preferably, yields the continuous process. For a start in this direction, see [3].

In order to illustrate the method in simple fashion, we shall consider a two-dimensional process,

$$\begin{aligned} x_1(n+1) &= x_1(n) - y_1(n) - r_1(n), \qquad x_1(0) = c_1, \\ x_2(n+1) &= g_2[x_1(n), x_2(n), y_2(n), r_2(n)], \qquad x_2(0) = c_2. \end{aligned} \tag{2}$$

The aim of the process is to choose $y_1(n)$ and $y_2(n)$, subject to constraints of the form

$$0 < a_1 \leq y_1(n) \leq a_2 , \qquad 0 \leq b_1 \leq y_2(n) \leq b_2 \tag{3}$$

so as to minimize the expected value of $[x_2(m) - x_0]^2$, where m is the "time" at which $x_1(m) = 0$. The $r_i(n)$ are independent random variables with given distributions.

The expected value is over the random variables r_1 and r_2 , where r_1 can depend upon the choice of y_1 and y_2 , but, in any case is subject to the condition that

$$y_1(n) + r_1(n) \geq a_3 > 0. \tag{4}$$

It follows that $x_1(n)$ is steadily decreasing as n increases.

The recurrence relation in (1) is valid until $x_1(n) = 0$. Properly, we should write

$$x_1(n + 1) = \max [0, x_1(n) - y_1(n) - r_1(n)]. \tag{5}$$

The process ends as soon as x_1 assumes the value zero.

3. Functional equations. It is clear that the minimum of the expected value of $[x_2(n) - x_0]^2$ depends upon c_1 and c_2 and only upon these variables assuming all other functions and distributions known and fixed. Let us then write

$$f(c_1 , c_2) = \min_{y_i} \exp_{r} [x_2(n) - x_0]^2. \tag{1}$$

We have

$$f(0, c_2) = (c_2 - x_0)^2, \tag{2}$$

and the principle of optimality, see [1], yields the functional equation

$$f(c_1 , c_2) = \min_{y_1, y_2} [\exp_{r_1, r_2} f(c_1 - y_1 - r_1 , g(c_1 , c_2 , y_2 , r_2)]. \tag{3}$$

There is no difficulty in treating the case in which the distribution of random effects depends upon the decisions that are made.

4. Probability of deviation. In place of mean-square deviation, let us consider the problem of determining y_1 and y_2 so as to minimize the probability that $| x_2 - x_0 | \geq d$.

As above, let

$$f(c_1 , c_2) = \min_{y_i} \text{prob } [| x_2 - x_0 | \geq d]. \tag{1}$$

Then

$$\begin{aligned} f(0, c_2) &= 1, \qquad | c_2 - x_0 | \geq d, \\ &= 0, \qquad | c_2 - x_0 | < d. \end{aligned} \tag{2}$$

while $f(c_1 , c_2)$ satisfies the same functional equation as in (3.3).

5. Discussion of computational solution. In order to determine the function $f(c_1, c_2)$ using a digital computer, we employ a discrete grid in (c_1 , c_2)-space. Let c_1 assume only the sequence of values $0, \delta, 2\delta, \cdots$, and c_2 a sequence of values $0, \Delta, 2\Delta, \cdots$. Since c_1 is monotonically decreasing as the process continues, we can use it as a "time" variable. Write

$$f(k\delta, c_2) \equiv f_k(c_2). \tag{1}$$

Then (3.3) may be written

$$f_k(c_2) = \min_{y_1, y_2 \quad r_1, r_2} [\exp f_p[g(k\delta, c_2, y_2, r_2)]], \tag{2}$$

where p is determined by the condition

$$p = [(c_1 - y_1 - r_1)/\delta], \tag{3}$$

the greatest integer contained in $(c_1 - y_1 - r_1)/\delta$.

Since $g(k\delta, c_2, y_2, r_2)$ in general will not be an integral multiple of Δ, we can either take as its value the nearest integer multiple of Δ, as we did in (3), or we can use interpolation, if more accurate results are desired.

The value of $f_0(c_2)$ is determined by the relation

$$f_0(c_2) = (c_2 - x_0)^2. \tag{4}$$

Consequently (2) furnishes a recurrence relation which enables us to compute the function $f_k(c_2)$ in terms of $f_n(c_2)$ for $n = 0, 1, \cdots, k - 1$. We thus have a feasible computational scheme.

6. Deterministic process. Returning to a purely deterministic process, as specified by (1.1), we may wish to determine y so that x is in some desired state at some subsequent time. One way of attacking this problem is to treat the problem of minimizing $[x_2(T) - x_0]^2$, where T is the first time at which $x_1(T)$ has its desired value. The functional equations are as above, without the averaging over the random behavior.

7. Linear equations and quadratic criteria. In general, the application of a straightforward functional equation approach is limited by dimensionality difficulties in the sense that functions of three or more variables cannot be readily stored in a fast memory. Consequently, the techniques described above must be aided and abetted by successive approximations of various types, a subject which has been discussed elsewhere. If, however, the guiding equations are linear, and the criteria function quadratic, then the sequence of functions $\{f_n(c)\}$ will consist of a sequence of quadratic functions in c. These functions are determined once the coefficients are determined. As we shall see, reasonably simple recurrence relations exist connecting the coefficients of $f_n(c)$ with those of $f_{n-1}(c)$.

Consider, to begin with, the problem of choosing the y_i so as to minimize the expected mean-square deviation

$$J_T(y) = \exp_r \left[(x(T) - a, x(T) - a) + \sum_{k=0}^{T} (y_k, By_k) \right]. \tag{1}$$

Here T assumes the values 0, 1, 2, $\cdots$, B is a positive definite matrix, a is a specified state vector, x and y are related by means of the linear relations

$$x_{n+1} = Ax_n + y_n + r_n, \qquad x_0 = c, \tag{2}$$

where $\{r_i\}$ is a set of independent, random vectors with identical distributions.

The process is assumed to proceed in the following fashion. We observe c, the initial state, and on this basis and the foregoing information, choose y_0, the initial control vector. Then a random effect r_0 occurs, yielding by way of (2) a new state vector $Ac + y_0 + r_0$. The process then continues in this way, stage-by-stage, a "feedback control" process.

Although this problem can be, and has been, treated by straightforward variational techniques, we shall treat it by functional equation methods. There is some merit in

doing this even in this case, and in addition we shall prepare the way for the following section devoted to a process of random duration.

Define the new sequence of functions $\{f_T(c)\}$ by means of the relation

$$f_T(c) = \min_{\{y\}} J_T(y). \tag{3}$$

Then

$$f_0(c) = (c - a, c - a), \tag{4}$$

and the principle of optimality yields the recurrence relation

$$f_n(c) = \min_{y_0} \exp_{r_0} [(y_0 , By_0) + f_{n-1}(Ac + y_0 + r_0)], \tag{5}$$

for $n = 1, 2, \cdots$.

Let us now show inductively that each $f_n(c)$ may be written in the form

$$f_n(c) = (c, M_n c) + 2(b_n , c) + u_n . \tag{6}$$

The result is obviously so for $n = 0$.

Substituting in (5), we have

$$\begin{aligned} f_n(c) = \min_{y_0} \exp_{r_0} [(y_0 , By_0) + [Ac + y_0 + r_0 , M_{n-1}(Ac + y_0 + r_0)] \\ + 2(b_{n-1} , Ac + y_0 + r_0) + u_{n-1}]. \end{aligned} \tag{7}$$

Taking expected values and using the result that

$$\min_y [(y, Cy) + 2(g, y)] = -(g, C^{-1}g), \tag{8}$$

whenever C is positive definite, we see that $f_n(c)$ has the form stated in (6). Carrying through the calculations, we obtain recurrence relations connecting M_n , b_n and d_n with M_{n-1} , b_{n-1} and d_{n-1} .

8. Linear process of random duration. Consider now a system specified by the equations

$$\begin{aligned} u_{n+1} &= u_n - r_{1n} , \qquad u_0 = c_0 , \\ x_{n+1} &= Ax_n + y_n + r_n , \qquad x_0 = c, \end{aligned} \tag{1}$$

where u_n and r_{1n} are scalars, x_n , y_n and r_n vectors. The process ends whenever u_n becomes zero or negative.

The quantity r_{1n} is a uniformly positive random variable, so that the process is always finite. The control vectors y_n are to be chosen so as to minimize the expected value of

$$J(y) = [x(m) - a, x(m) - a] + \sum_{k=0}^{m} (y_k , By_k), \tag{2}$$

where m is itself a random variable determined by the condition that it is the first integer for which u_m is negative or zero.

Write

$$f(c_0 , c) = \min_y \exp_r J(y). \tag{3}$$

Then

$$f(0, c) = (c - a, c - a), \tag{4}$$

and

$$f(c_0 , c) = \min_{y_0} \exp_{r_0} f(c_0 - r_{10} , Ac + y_0 + r_0). \tag{5}$$

Assume, as previously, that c_0 can assume only a discrete set of values with a similar condition on r_{10} . Let, suitably normalized, c_0 take the values 0, 1, $\cdots$, and r_{10} only the range of values $d_1 , d_1 + 1, \cdots , d_2$. Then, writing

$$f(k, c) \equiv f_k(c), \qquad k = 0, 1, 2, \cdots , \tag{6}$$

we may write (5) in the form

$$f_n(c) = \min_{y_0} \exp_{r_0} \left\{ \sum_{i=d_1}^{d_2} p_i f_{n-i}(Ac + y_0 + r_0) \right\} , \tag{7}$$

where

$$p_i = \text{the probability that } r_{10} = i. \tag{8}$$

The function $f_k(c)$ is identically zero for $k \leq 0$.

Once again, it is easy to see that each element of the sequence $\{f_k(c)\}$ is a quadratic function of c, of the form

$$f_k(c) = (c, M_k c) + 2(b_k , c) + u_k . \tag{9}$$

The recurrence relations connecting M_k , b_k , u_k with $M_{k-1} , b_{k-1} , u_{k-1}$ can be obtained from (7) in the way indicated above.

Bibliography

1. R. Bellman, *Dynamic programming*, Princeton University Press, Princeton, N. J., 1957
2. R. Bellman, I. Glicksberg and O. Gross, *On the "bang-bang" control problem*, Quart. Appl. Math. **14**, 11–18 (1956)
3. R. Bellman, *Functional equations in the theory of dynamic programming—VI: A direct convergence proof*, Ann. Math. **65**, 215–223 (1957)

DETERMINATION OF CHARACTERISTIC VALUES*

BY

MARK LOTKIN

Avco Mfg. Corp., Wilmington, Mass.

A method for computing the characteristic values of arbitrary matrices was recently proposed by the author in reference [1]. This method utilizes a sequence of unitary transformations which are designed to triangularize the original matrix, thereby producing the desired characteristic values along the diagonal of the triangularized form. Each of the unitary transformations is constructed in such a way as to reduce the norm of the upper-triangular part of the matrix. The basic function underlying this construction is a certain cubic polynomial whose coefficients depend upon the elements of the matrix to be reduced.

In this paper there is presented a refinement of the cubic polynomial, which is shown to possess properties that make it superior to the previous polynomial, on the basis of theoretical and practical considerations. Theoretically, the modified approach is seen to become identical with the Jacobi method for symmetric matrices, in certain cases; practically, the modification has been found to lead to more rapid convergence, at least for a considerable number of certain matrices that were subjected to both techniques.

A brief résumé of the relevant equations inherent in the procedure seems appropriate here. Let us assume, then, that the sequence of transformed matrices has reached the pth stage A_p, $p = 0, 1, 2, \cdots$, $A_0 = A$, and that it is consequently desired to construct $A_{p+1} = T_p^{-1} A_p T_p$, where the unitary matrix T_p has the "norm-reducing" property, *i.e.*, if

$$M_p = \sum_{\substack{r,s=1 \\ r<s}}^{n-1} | a_{rs}^{(p)} |^2$$

denotes the "upper-triangular" norm of A_p, and M_{p+1} denotes the corresponding quantity for A_{p+1}, then

$$M_{p+1} - M_p < 0. \tag{1}$$

Let the element $b = | b | \exp i\beta$, located in the (i, j) the position $i < j$, of A, be the "pivot" for the next transformation T_{p+1}, and let the elements in the position (i, i), (j, j), and (j, i) of A_p be denoted by $a = | a | \exp i\alpha$, d, and c, respectively; in general, let $a_{rs}^{(p)} = | a_{rs}^{(p)} | \exp (i\alpha_{rs}^{(p)})$.

It was shown in [1] that, for $R \neq 0$,

$$M_{p+1} - M_p < H(R, \theta) \equiv RF(R, \theta) \tag{2}$$

with

$$F(R, \theta) = C_3R^3 + C_2R^2 + C_1R + C_0 \tag{3}$$

---

*Received June 10, 1958; revised manuscript received November 28, 1958.

and

$$C_3 = |c|^2,$$

$$C_2 = -2|c|[|d|\cos(\theta+\gamma-\delta) - |a|\cos(\theta+\gamma-\alpha)],$$

$$C_1 = |d-a|^2 - 2|b||c|\cos(2\theta-\beta+\gamma) + \sum_{i+1\le k\le j-1}(|a_{ki}^{(p)}|^2 + |a_{jk}^{(p)}|^2),$$

$$C_0 = 2|b|[|d|\cos(\theta+\delta-\beta) - |a|\cos(\theta+\alpha-\beta)] + 2\sum_{i+1\le k\le j-1}[|a_{ik}^{(p)}||a_{jk}^{(p)}|\cos(\theta+\alpha_{jk}-\alpha_{ik}) - |a_{ki}^{(p)}||a_{kj}^{(p)}|\cos(\theta+\alpha_{ki}-\alpha_{kj})].$$

Any solution (R, θ) of the system

$$F(R, \theta) = 0 \tag{4}$$

$$\partial F/\partial\theta = 0 \tag{5}$$

then guarantees that $M_{p+1} < M_p$. Having determined R, θ, the elements of the transformation matrix T_p are found by means of

$$r = (R^2+1)^{-1/2}, \operatorname{sgn} r = \operatorname{sgn} R, t = r\exp i\theta, \tag{6}$$

and the new elements of A_{p+1} are determined from the relationships (8) through (15) of reference [1].

In addition to the relationships (16) between the elements of A_{p+1} and A_p, stated in [1], there exist further identities between these elements, of value in the actual performance of numerical calculations; some of these are exhibited below.

A short calculation shows, for example, that

$$|a_1|^2 + |b_1|^2 + |c_1|^2 + |d_1|^2 = |a|^2 + |b|^2 + |c|^2 + |d|^2, \tag{7}$$

by virtue of $r^2(1+R^2) = 1$. For the same reason,

$$|a_{ik}^{(1)}|^2 + |a_{jk}^{(1)}|^2 = |a_{ik}|^2 + |a_{jk}|^2 \tag{8}$$

$$|a_{ki}^{(1)}|^2 + |a_{kj}^{(1)}|^2 = |a_{ki}|^2 + |a_{kj}|^2 \tag{9}$$

for $1 \le k \le n$, $k \ne i, j$.

Further, whenever $\theta = 0$,

$$b_1 - c_1 = b - c. \tag{10}$$

As stated in the introductory paragraph, the determination of R, θ was previously based on Eqs. (4) and (5).

Now a possibly larger decrease in the norm M than indicated by (4) and (5) may be obtained by choosing R such that

$$M_1 - M \le \min_{-\infty<R<\infty} H(R, \theta).$$

The values of R at which min H occurs must then satisfy

$$\partial H/\partial R \equiv 4C_3R^3 + 3C_2R^2 + 2C_1R + C_0 = 0 \tag{11}$$

$$\partial H/\partial\theta = 0. \tag{12}$$

If $R = R_m$ satisfies (11), then clearly

$$M - M_1 \geq R_m^2(3C_3R_m^2 + 2C_2R_m + C_1).$$

Now for fixed θ, $H(0, \theta) = 0$, $(\partial H/\partial R)_{0,\theta} = 0$, $(\partial^2 H/\partial R^2)_{0,\theta} = 2\,C_1$. Thus min $H < 0$ whenever $C_0 \neq 0$. If $C_0 = 0$, but $C_1 < 0$, then still min $H < 0$. Only if $C_0 = 0$, $C_2^2 - 4C_1C_3 \leq 0$, is min H equal to zero. Thus in general (11) may be considered superior to (4) for the reduction of the super-diagonal norm. If Eq. (11) has, for fixed θ, two negative minima, then we choose for the transformation $z = R \exp i\theta$ that root R for which min min H is assumed.

The superiority of (11) may be deduced also from the study of certain second order matrices, which are obviously of basic importance in this problem.

I. Let us consider

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{13}$$

with $a = d$. If A is skew-hermitian, then $a = d = 0$. It may be assumed that $bc \neq 0$. Since now $\alpha = \delta$, (11) becomes

$$4 \mid c \mid R(\mid c \mid R^2 - \mid b \mid) = 0,$$

so that $R = \mid b/c \mid^{1/2}$, and min $H(R, \theta) = - \mid b \mid^2$. With $z = \mid b/c \mid^{1/2} \exp 2^{-1}(\beta - \gamma)i$, and $r^2 = \mid c \mid/(\mid b \mid + \mid c \mid)$, the transformation equations lead to

$$a_1 = a + (bc)^{1/2}$$

$$d_1 = a - (bc)^{1/2}$$

$$b_1 = 0$$

$$c_1 = (b/\mid b \mid)(\mid c \mid - \mid b \mid).$$

The expressions for a_1, d_1 are, naturally, the exact roots for the matrix (13), with $a = d$. Thus $M_1 - M = - \mid b^2$.

II. Again let us consider matrix A as defined by (13), now subject to the following conditions:

(i) $\alpha = \delta$

(ii) $\beta + \gamma = 2\alpha$

(iii) $\theta = 2^{-1}(\beta - \gamma)$.

Then

$$C_3 = \mid c \mid^2$$

$$C_2 = -2 \mid c \mid (\mid d \mid - \mid a \mid)$$

$$C_1 = (\mid d \mid - \mid a \mid)^2 - 2 \mid bc \mid$$

$$C_0 = 2 \mid b \mid (\mid d \mid - \mid a \mid),$$

whence

$$S(R) = [\mid c \mid R^2 - (\mid d \mid - \mid a \mid)R - \mid b \mid][2 \mid c \mid R - (\mid d \mid - \mid a \mid)] \tag{14}$$

$$H(R) = R[|\, c \,| R^2 - (|\, d \,| - |\, a \,|)R - 2 \,|\, b \,|][|\, c \,| R - (|\, d \,| - |\, a \,|)]. \tag{15}$$

If $|\, c \,| R^2 - (|\, d \,| - |\, a \,|)\, R - |\, b \,| = 0$, then $H(R) = -|\, b \,|^2$. If, however, $2 |\, c \,| R - (|\, d \,| - |\, a \,|) = 0$, then $H(R) = |\, c \,| R^2 (|\, c \,| R^2 + 2 |\, b \,|) > 0$. Consequently, $\min H(R) = -|\, b \,|^2$ is assumed at the roots of $|\, c \,| R^2 - (|\, d \,| - |\, a \,|)R - |\, b \,| = 0$. Therefore, $b_1 = 0$, so that the characteristic values of A appear immediately as a_1, d_1.

III. Now let the nth order matrix A be hermitian, i.e., $a_{ij} = \bar{a}_{ij}$, a_{ii} real.

Then it is seen that the cubic polynomial $F(R, \theta)$ of (2) becomes

$$\begin{aligned} C_3 &= |\, b \,|^2 \\ C_2 &= -2 |\, b \,| [|\, d \,| \cos(\theta + \gamma - \delta) - |\, a \,| \cos(\theta + \gamma - \alpha)] \\ C_1 &= |\, d - a \,|^2 - 2 |\, b \,|^2 \cos(2\theta - \beta + \gamma) \\ C_0 &= 2 |\, b \,| [|\, d \,| \cos(\theta + \delta - \beta) - |\, a \,| \cos(\theta + \alpha - \beta)]. \end{aligned}$$

for the choice of $\theta = 2^{-1}(\beta - \gamma)$ above expressions become

$$\begin{aligned} C_3 &= |\, b \,|^2 \\ C_2 &= -2 |\, b \,| (d - a) \\ C_1 &= (d - a)^2 - 2 |\, b \,|^2 \\ C_0 &= 2 |\, b \,| (d - a), \end{aligned}$$

and, consequently,

$$H(R, \theta) \equiv RF(R, \theta) = R[|\, b \,| R^2 - (d - a)R - 2 |\, b \,|][|\, b \,| R - (d - a)]$$

$$S(R, \theta) \equiv \partial H / \partial R = 2[|\, b \,| R^2 - (d - a)R - |\, b \,|][2 |\, b \,| R - (d - a)].$$

Thus again $H(R) = -|\, b \,|^2$ for the roots R of the quadratic factor of $S(R, \theta)$. This, naturally, implies again that $b_1 = 0$.

IV. The matrix

$$B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \tag{16}$$

has been mentioned as one which defies direct treatment by Greenstadt's method [2], as well as by the method of reference [1]. However, it is stated in [3] that by applying a transformation to B which effects a rotation through $\theta = \pi/4$, the transformed matrix becomes tractable by Greenstadt's method, and that in twelve cyclically executed annihilations of the respective pivot the matrix B becomes triangularized to a sufficient degree of accuracy.

It will be seen that the same preparatory rotation through $\theta_0 = \pi/4$, followed by two transforms of the type determined by (11), exactly diagonalizes the matrix (16).

Since the transpose B^T of B has the lower superdiagonal norm, we subject B^T rather than B to the sequence of unitary transformations. The preliminary rotation is effected by

$$T = \begin{bmatrix} \cos\theta_0 & 0 & -\sin\theta_0 \\ 0 & 1 & 0 \\ \sin\theta_0 & 0 & \cos\theta_0 \end{bmatrix}, \tag{17}$$

leading to

$$B_1 \equiv T^{-1}B^TT = \begin{bmatrix} 3/2 & 2^{-1/2} & 1/2 \\ 2^{-1/2} & 1 & -2^{-1/2} \\ -1/2 & 2^{-1/2} & 1/2 \end{bmatrix}.$$

Let us choose next the element $a_{12}^{(1)} = 2^{-1/2}$ as the pivot b. Then case II discussed previously is found to apply. With $z = 2^{-1/2}$ one obtains

$$B_2 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1/2 & -3^{1/2}/2 \\ 0 & 3^{1/2}/2 & 1/2 \end{bmatrix}.$$

The superdiagonal norm $M(B_1) = 5/4$ has thus been reduced by the amount $b^2 = 2^{-1}$ to $M(B_2) = 3/4$.

Next we take $a_{23}^{(2)} = -3^{1/2}/2$ as the pivot b. Here case I obtains. Therefore, $R = 1$, $z = i$, and

$$B_3 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & (1/2)(1 + i/3^{1/2}) & 0 \\ 0 & 0 & (1/2)(1 - i/3^{1/2}) \end{bmatrix}, \tag{18}$$

so that B has actually been reduced to diagonal form.

The diagonalization of B^T has thus been achieved by subjecting it to the unitary transformation $B_3 = P^{-1}B^TP$, with

$$P = \begin{bmatrix} 2^{-1/2} & 0 & -2^{-1/2} \\ 0 & 1 & 0 \\ 2^{-1/2} & 0 & 2^{-1/2} \end{bmatrix} \begin{bmatrix} (2/3)^{1/2} & -3^{-1/2} & 0 \\ 3^{-1/2} & (2/3)^{1/2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2^{-1/2}i & -2^{-1/2} \\ 0 & 2^{-1/2} & -2^{-1/2}i \end{bmatrix}$$

$$= \begin{bmatrix} 3^{-1/2} & -(1/2) - (2\cdot 3^{1/2})^{-1}i & (2\cdot 3^{1/2})^{-1} + i/2 \\ 3^{-1/2} & 3^{-1/2}i & -3^{-1/2} \\ 3^{-1/2} & (1/2) - (2\cdot 3^{1/2})^{-1}i & (2\cdot 3^{1/2})^{-1} - i/2 \end{bmatrix}. \tag{19}$$

According to a theorem of Toeplitz (see, e.g., [4]) a matrix M can be reduced by unitary transformations to the diagonal form if and only if the matrix M is normal: $M^{c^T}M = MM^{c^T}$, where M^{c^T} denotes the conjugate transpose of M. Thus the matrix B is seen to be normal. The interesting question then arises whether a class of normal matrices of which B is a member can be reduced to diagonal form by the general technique of this paper. This question can be answered in the affirmative; the results will be published elsewhere.

V. While the choice of a particular value of θ may be appropriate, in special situations, in general the condition (12) may have to be considered, for optimum results. In the example discussed here the values of $\theta = 0, \pi/6, \pi/3, \cdots 5\pi/6$ were applied to each pivot, which was always chosen to be the element of largest modulus. For the transformation $z = R \exp i\theta$ that pair (R, θ) was selected for which min $H(R, \theta)$ is assumed.

The following matrix is taken from [1]:

$$A = \begin{bmatrix} 1 & 0 & -2 \\ 2 & -1 & 2 \\ 2 & 1 & 0 \end{bmatrix}; \tag{20}$$

its characteristic values are -2, $1 \pm 2i$. Using the method of [1], eight iterations of A produced the diagonal terms shown under (33) of reference [1], viz.

$$-1.97139 - .07432i, \qquad .95014 + 2.11992i, \qquad 1.02125 - 2.04650i.$$

The use of the refinement based on Eq. (11) led to

$$-1.99489 + .01100i, \qquad .99161 + 2.00158i, \qquad 1.00328 - 2.01258i,$$

clearly a considerably improved result over the previous one. The super-diagonal norm at this stage had been reduced from 8.000 to 1.424×10^{-3}.

VI. The matrix

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} & \frac{1}{10} \\ \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} & \frac{1}{10} & \frac{1}{11} \end{bmatrix} \tag{21}$$

is one of a sequence of non-symmetric matrices of extremely bad "condition" [5]. Matrix C is nearly singular; the absolute value of its determinant is $(31\ 05\ 2236\ 7232 \cdot 10^{-5})^{-1} = .3220\ 3799 \cdot 10^{-16}$. It is well known—see, e.g., Todd [6]—that certain calculations with these matrices, such as inversions, determination of characteristic values, etc., suffer from "numerical instability". For machines with twelve decimals, employing floating point arithmetic, attempts to invert matrices of even the eighth order have been doomed to failure.

The characteristic values λ_n of C, calculated by means of a determinantal method, and arranged in order of decreasing magnitude, are listed in Table 1.

TABLE 1.

Characteristic values of C.

n	λ_n
1	$.2132\ 3763 \times 10^{1}$
2	$-.2214\ 0681 \times 10^{0}$
3	$-.3184\ 3305 \times 10^{-1}$
4	$-.8983\ 2330 \times 10^{-3}$
5	$-.1706\ 2788 \times 10^{-4}$
6	$-.1397\ 4990 \times 10^{-6}$

It is seen from this table that $\lambda_1 \cdot \lambda_2 \cdots \lambda_6 = .3220\ 3799 \times 10^{-16} = \det C$, to eight significant figures.

The reduction technique described above was programmed for the IBM 704 machine. Single precision arithmetic, and floating point with eight significant figures was used. At each step, the super-diagonal element of largest absolute value was taken as the pivot $b = a_{i,i}$ of the next transformation. Some of the results are shown in Table 2.

TABLE 2.

Triangularization of Matrix C

p	M_p	$a_{ii}^{(M)}$	$a_{jj}^{(m)}$
0	$.494 \times 10^{0}$	1.0000 0000	$.9090\ 9090 \times 10^{-1}$
10	$.236 \times 10^{-2}$	$.2124\ 5647 \times 10^{1}$	$.1250\ 4932 \times 10^{-2}$
20	$.553 \times 10^{-4}$	$.2130\ 8000 \times 10^{1}$	$.2123\ 4414 \times 10^{-3}$
30	$.296 \times 10^{-5}$	$.2132\ 2888 \times 10^{1}$	$-.6031\ 1972 \times 10^{-4}$
40	$.385 \times 10^{-6}$	$.2132\ 3244 \times 10^{1}$	$-.7919\ 3310 \times 10^{-4}$
50	$.616 \times 10^{-7}$	$.2132\ 3678 \times 10^{1}$	$-.2023\ 3858 \times 10^{-5}$
60	$.544 \times 10^{-8}$	$.2132\ 3763 \times 10^{1}$	$-.2646\ 2287 \times 10^{-5}$
70	$.980 \times 10^{-9}$	$.2132\ 3705 \times 10^{1}$	$-.1103\ 9584 \times 10^{-5}$
80	$.149 \times 10^{-9}$	$.2132\ 3751 \times 10^{1}$	$-.1097\ 4208 \times 10^{-5}$
90	$.239 \times 10^{-10}$	$.2132\ 3772 \times 10^{1}$	$-.1057\ 6060 \times 10^{-5}$
100	$.448 \times 10^{-11}$	$.2132\ 3763 \times 10^{1}$	$-.1451\ 8099 \times 10^{-6}$
110	$.647 \times 10^{-12}$	$.2132\ 3765 \times 10^{1}$	$-.1315\ 6736 \times 10^{-6}$
120	$.143 \times 10^{-12}$	$.2132\ 3766 \times 10^{1}$	$-.1652\ 8475 \times 10^{-6}$
130	$.245 \times 10^{-13}$	$.2132\ 3765 \times 10^{1}$	$-.1634\ 9713 \times 10^{-6}$
140	$.581 \times 10^{-14}$	$.2132\ 3765 \times 10^{1}$	$-.1371\ 2160 \times 10^{-6}$
150	$.834 \times 10^{-15}$	$.2132\ 3765 \times 10^{1}$	$-.1409\ 3837 \times 10^{-6}$

In this table p denotes the number of iterations, M_p the super-diagonal norm of A_p, $a_{ii}^{(M)}$, $a_{jj}^{(m)}$ the diagonal elements in A_p of largest and smallest absolute value, respectively.

The diagonal elements at $p = 150$, arranged in order of decreasing magnitude, are:

$$.2132\ 3765 \times 10^{1}$$

$$-.2214\ 0677 \times 10^{0}$$

$$-.3184\ 3361 \times 10^{-1}$$

$$-.8983\ 2775 \times 10^{-3}$$

$$-.1705\ 5897 \times 10^{-4}$$

$$-.1409\ 3837 \times 10^{-6}$$

Thus the dominant characteristic value is determined at this stage to about seven significant figures, while the smallest one is known to about three. The rate of decrease of M_p, from $.5 \times 10^{0}$ to $.8 \times 10^{-15}$, would seem to indicate a "linear" type of convergence. It is obvious that further improvements of the results will be achieved once a number of basic routines that are presently in the program have been sharpened. Such routines are concerned with the conversion of numbers from the decimal to binary system, the calculation of trigonometric functions, the location of roots of polynomials, and other operations required in the method.

VII. Among the many other matrices that have been reduced satisfactorily we mention here the Hilbert matrices. These are symmetric matrices whose elements are

$a_{ij} = (i + j - 1)^{-1}$, $i, j = 1, 2, 3, \cdots$. Some results for the eighth order matrix are exhibited in Table 3.

TABLE 3.

Characteristic values of a Hilbert matrix

p	M_p	$a_{ii}^{(M)}$	$a_{ii}^{(m)}$
0	$.882 \times 10^{0}$	1.0000 0000	$.6666\ 6667 \times 10^{-1}$
10	$.262 \times 10^{-1}$	$.1693\ 0662 \times 10^{1}$	$.4273\ 0205 \times 10^{-2}$
20	$.388 \times 10^{-3}$	$.1695\ 9118 \times 10^{1}$	$.4030\ 7163 \times 10^{-2}$
30	$.543 \times 10^{-5}$	$.1695\ 9389 \times 10^{1}$	$.3309\ 0313 \times 10^{-4}$
40	$.111 \times 10^{-6}$	$.1695\ 9390 \times 10^{1}$	$.1518\ 9860 \times 10^{-4}$
50	$.123 \times 10^{-8}$	$.1695\ 9391 \times 10^{1}$	$.6202\ 6355 \times 10^{-5}$
60	$.211 \times 10^{-11}$	$.1695\ 9391 \times 10^{1}$	$.3315\ 4293 \times 10^{-6}$
70	$.364 \times 10^{-13}$	$.1695\ 9391 \times 10^{1}$	$.7117\ 3257 \times 10^{-8}$
75	$.424 \times 10^{-16}$	$.1695\ 9391 \times 10^{1}$	$.6802\ 6828 \times 10^{-8}$

The intermediate diagonal elements in A_{75} are:

$$.2981\ 2524 \times 10^{0}$$

$$.2621\ 2851 \times 10^{-1}$$

$$.1467\ 6944 \times 10^{-2}$$

$$.5437\ 2030 \times 10^{-4}$$

$$.1297\ 1307 \times 10^{-5}$$

$$.1589\ 9581 \times 10^{-7}$$

The trace of the matrix, which is theoretically equal to the sum of the characteristic values, is 2.0218 0042. The sum of the diagonal elements at $p = 75$, on the other hand, is found to be 2.0218 006. The well-known Givens method for the characteristic values of real symmetric matrices results in a corresponding value of 2.0218 002.

References

1. M. Lotkin, *Characteristic values of arbitrary matrices*, Quart. Appl. Math. XIV, 267-275 (1956).
2. J. Greenstadt, *A method for finding roots of arbitrary matrices*, MTAC 9, 47-52 (1955).
3. R. L. Causey, *Computing eigenvalues of non-hermitian matrices by methods of Jacobi type*, J. Soc. Ind. Appl. Math. 6, 172-181 (1958).
4. C. C. MacDuffee, *The theory of matrices*, Chelsea, New York, 1956, p. 76
5. M. Lotkin, *A set of test matrices*, MTAC 9, 153-161 (1955)
6. J. Todd, *The condition of the finite segments of the Hilbert matrix*, NBS Appl. Math. Ser. 39, 1954

SOME PROPERTIES OF RATIONAL TRANSFER FUNCTIONS AND THEIR LAPLACE TRANSFORMATIONS*

BY

ARMEN H. ZEMANIAN

College of Engineering, New York University

Introduction. The behavior of a fixed, linear system at a pair of output terminals due to a signal impressed at a pair of input terminals is determined by the transfer function between these two terminals. More precisely, this transfer function $Z(s)$ is the ratio of the Laplace transform of the output function of time to the Laplace transform of the input function of time. Moreover, if the system is lumped and finite, the transfer function will be a rational function of the complex frequency variable $s = \sigma + j\omega$ and may be written as the ratio of two polynomials in s

$$Z(s) = K \frac{s^n + a_{n-1}s^{n-1} + \cdots + a_0}{s^m + b_{m-1}s^{m-1} + \cdots + b_0} = K \frac{N(s)}{D(s)}. \tag{1}$$

The constant multiplier K and the coefficients a_i and b_i are all real numbers. It will be assumed still further that the system is a stable one. That is, if the input signal is removed, the output signal becomes arbitrarily small with increasing time. This restricts the denominator $D(s)$ to being a Hurwitz polynomial (a polynomial whose roots all have negative real parts). Finally, the transfer function of any physically realizable system must become arbitrarily small as the magnitude of the complex frequency s becomes arbitrarily large. This is due to the stray capacities and parameter dissipation that must exist in such a system. Thus, this discussion will be restricted, in addition, to those transfer functions that have more poles than zeros ($m > n$). This insures that the Fourier transform, relating the transfer function $Z(j\omega)$ to its corresponding unit impulse response $W(t)$, exists. Only those transfer functions which satisfy the aforementioned conditions will be considered in this paper. Therefore, all unit impulse responses will be a finite sum of terms of the form,

$$At^{\mu}e^{-\alpha t} \cos(\beta t + \gamma),$$

where A, α, β, and γ are constants with $\alpha > 0$ and μ is either zero or a positive integer.

In a previous paper [1], certain classes of these transfer functions have been investigated and many restrictions on the transient responses corresponding to such transfer functions were shown to exist. These classes are related to the class of positive real functions for which $m - n$ equals one. However, the sum of all these classes does not include every transfer function of the type considered here. A transfer function that belongs to one of these classes is called a class k function and the integer k identifies the particular class. Throughout this paper, the positive integer k will represent the number of poles in excess of the number of zeros.

The purpose of this paper is to determine other properties of the class k functions and then to extend the restrictions on their corresponding transient responses to the

*Received June 25, 1958.

transient responses of any transfer function. That is, a means will be developed for determining bounds on the transient responses of any transfer function, whether or not that transfer function is a class k function, by relating it to a class k function.

The class k functions. For this investigation, the form of the definition of class k functions will be altered somewhat from that given in [1]. It should be emphasized, however, that these two definitions are the same in that they define exactly the same classes of functions. The new definition leads to simpler expressions in this paper.

Consider the function $Z_q(s)$ which is obtained by successively integrating the transfer function $Z(s)$ according to (2)

$$Z_q(s) = \int_s^\infty d\,s_{q-1} \int_{s_{q-1}}^\infty d\,s_{q-2} \cdots \int_{s_1}^\infty Z(s_0)\,d\,s_0\,. \tag{2}$$

The integer q is restricted to being less than or equal to the integer $k - 1$. The right-hand side of (2) is, in general, a multivalued function and, if all the poles of $Z(s)$ are simple, it may be written as

$$(-1)^q K \sum_{i=1}^{m} k_i \frac{(s + \rho_i)^{q-1}}{(q-1)!}\left[\ln (s + \rho_i) - \sum_{p=1}^{q-1} \frac{1}{p}\right]$$

where the ρ_i are the poles of $Z(s)$ and the k_i are the residues of these poles. [$Z_q(s)$ will be single-valued when $Z(s)$ has only one pole of multiplicity m.] For the purposes of this paper, only one branch of this multivalued function is needed. Therefore, given any $Z(s)$ having at least two more poles than zeros, branch cuts consisting of straight line segments connecting all the poles will be assumed. In addition, these straight line segments will be so chosen that any pair of poles can be connected by a path traced exclusively along these segments. That is, it will be assumed that these branch cuts form a tree with the distinct poles at its vertices. The number of these branch cuts will be one less than the number of distinct poles of $Z(s)$. Finally, it will be assumed throughout this paper that the paths of integration in (2) never cross any of these branch cuts. Under these conditions, (2) defines a single-valued function $Z_q(s)$.

Representing the real and imaginary parts of s_q and $Z_q(j\omega)$ by $s_q = \sigma_q + j\omega_q$ and $Z_q(j\omega) = R_q(\omega) + jI_q(\omega)$, the real part of $Z_q(j\omega)$ is given by the following expressions. For q odd,

$$R_q(\omega) = (-1)^{(q+1)/2} \int_\omega^\infty d\omega_{q-1} \int_{\omega_{q-1}}^\infty d\omega_{q-2} \cdots \int_{\omega_1}^\infty I(\omega_0)\,d\omega_0 \tag{3}$$

and for q even,

$$R_q(\omega) = (-1)^{q/2} \int_\omega^\infty d\omega_{q-1} \int_{\omega_{q-1}}^\infty d\omega_{q-2} \cdots \int_{\omega_1}^\infty R(\omega_0)\,d\omega_0\,, \tag{4}$$

where $R(\omega)$ and $I(\omega)$ are the real and imaginary parts of $Z(j\omega)$, respectively.

The class k functions may now be defined as follows.

Definition. The transfer function $Z(s)$ will be called a class k function, where $k = m - n$, if $Z_{k-1}(s)$ is a positive real function in the half plane $\sigma \geq 0$.

An immediate consequence of this definition is easily obtained. The necessary and sufficient condition for $Z(s)$ to be a class k function is that $R_{k-1}(\omega) \geq 0$ for all ω. This follows from the definition since $Z_{k-1}(s)$ is an analytic function and, moreover, $Z_{k-1}(\sigma)$ is a real function for $\sigma \geq 0$.

Furthermore, class k functions may be generated from the class 1 functions by differentiation. More generally, if $Z(s)$ is class k, then $(-1)^h d^h Z/ds^h$ is class $(k+h)$. Another elementary property of these functions is that $\alpha Z_a(s) + \beta Z_b(s)$ is class k if $Z_a(s)$ and $Z_b(s)$ are each at least class k and α and β are positive numbers.

Realizability conditions on the impulse response. It will be assumed throughout this paper that all driving functions of time are applied at $t = 0$. The unit impulse response $W(t)$ is, therefore, related to the transfer function $Z(s)$ through the Laplace transform,

$$Z(s) = \int_0^\infty W(t)e^{-st}\,dt, \tag{5}$$

where σ is greater than the real part of any pole of $Z(s)$. Moreover, the function $Z_q(s)$ may be obtained by integrating (5) under the integral sign q times. The resulting expression (6) is valid for $q \leq k-1$ [2, Theorem 17, p. 273].

$$Z_q(s) = \int_0^\infty \frac{W(t)}{t^q}\,e^{-st}\,dt. \tag{6}$$

A well known result of circuit theory is that, if $Z(s)$ is a positive real function, then it is expressible in terms of three positive definite quadratic forms which are related to the total stored energy and power dissipation in the network [3, Chap. 4]. There is a somewhat similar result in the theory of Laplace transform which will place restrictions on the unit impulse response corresponding to a positive real $Z(s)$ [4, Theorem 6]. More precisely, $Z(s)$ is a class 1 function if and only if $W_e(t)$ is a positive definite function where $W_e(t)$ is the even function given by

$$W_e(t) = \begin{cases} W(t) & \text{for } t \geq 0, \\ W(-t) & \text{for } t \leq 0. \end{cases}$$

By definition, $W_e(t)$ is positive definite if the Hermitian form

$$\sum_{i=1}^{n}\sum_{j=1}^{n} W_e(t_i - t_j)x_i x_j \tag{7}$$

is non-negative for all values of the real numbers $t_1, t_2, \cdots, t_n$ and $x_1, x_2, \cdots, x_n$ and for all n.

Applying these known results to expression (6), the following realizability theorem is immediately obtained.

Theorem 1. *The necessary and sufficient condition for $Z(s)$ to be a class k function is that $W_e(t)/|t|^{k-1}$ be a positive definite function. That is, $Z(s)$ is class k if and only if*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{W_e(t_i - t_j)}{|t_i - t_j|^{k-1}}\,x_i x_j \geq 0 \tag{8}$$

for all values of real numbers $t_1, t_2, \cdots t_n$ and $x_1, x_2, \cdots, x_n$ and for all n.

There are several necessary and sufficient conditions for a function to be positive definite [4 through 8] which could similarly be used for realizability conditions. However, all of these are, in general, too complicated to be useful as a practical test.

A condition which is sufficient (though not necessary) and at the same time quite

simple to apply may be obtained by applying Theorem 5 of [9] to the function $W(t)/t^{k-1}$ for the case where the integer m, defined in that theorem, equals one. This yields the following sufficiency test for a class k function.

Theorem 2. If the second derivative of $W(t)/t^{k-1}$ is non-negative for $t > 0$, then $Z(s)$ will be a class k function.

A necessary but not sufficient condition on the impulse response, in order that it be the Laplace transform of a class k function was given in [1, Theorem 9]. A new proof, which is considerably simpler than the one given previously, may be constructed by using an operation transform pair of the Laplace transform.

Theorem 3. If $Z(s)$ is class k, then, for $t \geq 0$,

$$| W(t) | \leq \frac{Kt^{k-1}}{(k-1)!}. \tag{9}$$

Proof. By the definition of class k functions, the function $Z_{k-1}(s)$ is positive real and the first term of its inverse series expansion is

$$\frac{K}{(k-1)!s}, \tag{10}$$

where $k = m - n$. It has been shown that such a positive real function has a unit impulse response which is bounded by the coefficient of the first term in its inverse series expansion [9, Theorem 1]. Moreover, each integration in expression (2) corresponds to the division of $W(t)$ by the time variable t. This transformation of the operations of complex integration and division by t is a well-known property of the Laplace transformation [2, Theorem 17, p. 273]. Thus, for $t \geq 0$,

$$\left| \frac{W(t)}{t^{k-1}} \right| \leq \frac{K}{(k-1)!},$$

which is the desired conclusion.

By use of the superposition integral, the previous result may be extended to those transfer functions that are products of class k functions. This is applicable, for instance, to a system of cascaded amplifiers each of whose interstage networks is class k. It may also be applied to the transfer functions of ladder networks [10].

Theorem 4. If the functions $A(s)$, $B(s)$, $\cdots$, $H(s)$ are class a, class b, $\cdots$, and class h, respectively, and their constant multipliers are K_a, K_b, $\cdots$, and K_h, respectively, then the unit impulse response $W(t)$ corresponding to the product of these functions, $A(s)\ B(s)\ \cdots\ H(s)$, is bounded by

$$| W(t) | \leq \frac{K_a K_b \cdots K_h}{(a+b+\cdots+h-1)!}\, t^{(a+b+\cdots+h-1)} \tag{11}$$

for $t \geq 0$.

Proof. Let $W_{ab}(t)$ be the inverse Laplace transform of the product $A(s)B(s)$. By the superposition integral,

$$| W_{ab}(t) | \leq \int_0^t | W_a(\tau) |\, | W_b(t-\tau) |\, d\tau.$$

But, since $A(s)$ is class a and $B(s)$ is class b,

$$| W_{ab}(t) | \leq \frac{K_a K_b}{(a-1)!(b-1)!} \int_0^t \tau^{a-1}(t-\tau)^{b-1}\, d\tau.$$

Integrating this expression by parts $b-1$ times,

$$| W_{ab}(t) | \leq \frac{K_a K_b}{(a+b-2)!} \int_0^t \tau^{a+b-2}\, d\tau = \frac{K_a K_b}{(a+b-1)!}\, t^{(a+b-1)}.$$

Applying this process repeatedly to the sequence of products $A(s)B(s)$, $A(s)B(s)C(s)$, $\cdots$, $A(s)B(s) \cdots H(s)$, expression (11) will be obtained.

A relationship between an arbitrary transfer function and the class k functions. One of the objects of this paper is to develop a method of determining bounds on the transient responses of any transfer function satisfying the restrictions enumerated in the introduction. This will be accomplished by relating such a transfer function to a class k function as follows.

Theorem 5. *For any given $Z(s)$, a real number c can be found such that $Z(s+c)$ is class k. Moreover, if $Z(s+c)$ is class k, then $Z(s+d)$ is class k for all d greater than c.*

Proof. By the definition of class k functions, $Z(s)$ will be class k only if $Z_{k-1}(s)$ is a positive real function. Now consider the loci of Re $[Z_{k-1}(s)] = 0$. Since $Z_{k-1}(s)$ behaves as $1/s$ in the neighborhood of $s = \infty$, only one such locus exists in a sufficiently small neighborhood of $s = \infty$. Moreover, this locus is asymptotic to a line parallel to the imaginary axis. Thus, for c sufficiently large, the line $s = c + j\omega$, which is parallel to the imaginary axis, will not intersect any locus of Re $[Z_{k-1}(s)] = 0$. Furthermore, Re $[1/s] \geq 0$ for $\sigma \geq 0$. Therefore, Re $[Z_{k-1}(s)] > 0$ for $\sigma \geq c$.

The second statement of this theorem follows from the fact that a positive real function of a positive real function is positive real.

Theorem 6. *If the real number c is such that $Z(s+c)$ is class k, then c cannot be less than $(a_{n-1} - b_{m-1})/k$, where a_{n-1} and b_{m-1} are the coefficients indicated in* (1). *Moreover, c cannot be less than the real parts of all the poles of $Z(s)$.*

Proof. Integrating the inverse series expansion of $Z(s)$ according to (2), a series expansion for $Z_{k-1}(s)$ may be found

$$Z_{k-1}(s) = \frac{K}{k!}\left(\frac{k}{s} + \frac{a_{n-1} - b_{m-1}}{s^2} + \cdots\right).$$

It can be seen from the first two terms that a locus of Re $[Z_{k-1}(s)] = 0$ is asymptotic to the line,

$$s = \frac{a_{n-1} - b_{m-1}}{k} + j\omega,$$

for ω sufficiently large. Moreover, such loci are continuous curves which pass through the singularities of $Z_{k-1}(s)$. Since all the singularities of a positive real function are excluded from the right half s plane, the constant c cannot be less than $(a_{n-1} - b_{m-1})/k$ if the line $s = c + j\omega$ is not to intersect a locus of Re $[Z_{k-1}(s)] = 0$ and if $Z_{k-1}(s+c)$ is to be analytic in the right half s plane. Moreover, the singularities of $Z_{k-1}(s)$ occur at the same points as the poles of $Z(s)$ so that c cannot be less than the real part of any such pole.

One physical interpretation of the addition of a positive constant c to the complex frequency variable s is that dissipation is added uniformly to all the reactive elements of any network which realizes $Z(s)$. In particular, every inductance L_i has a resistance $R_i = L_i c$ inserted in series with it and every condenser C_i has a conductance $G_i = C_i c$ connected in parallel with it [3, pp. 706–708].

A class k function $Z(s + c)$ can be obtained from a given $Z(s)$ in the following way. Consider the polynomial in the numerator of $Z(s)$

$$N(s) = s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0 . \tag{12}$$

The polynomial $N(s + c)$ may be determined by setting μ equal to $s + c$ and expanding $N(\mu)$ into a Taylor series around the point $\mu = c$

$$N(s + c) = s^n + \frac{s^{n-1}}{(n-1)!}\cdot\frac{d^{n-1}N(s)}{ds^{n-1}}\Bigg]_{s=c} + \cdots + s\frac{dN(s)}{ds}\Bigg]_{s=c} + N(c). \tag{13}$$

A simple method exists for obtaining the coefficients of $N(s + c)$ from the coefficients of $N(s)$ without having to perform the differentiations indicated in (13) [11, pp. 52–54]. This procedure is especially useful when a numerical value for c is used. Divide $N(s)$ by $s - c$ until a constant remainder r_0 is obtained.

$$\frac{N(s)}{s-c} = Q_{n-1}(s) + \frac{r_0}{s-c}. \tag{14}$$

The value of $N(c)$ is r_0 . Dividing the quotient $Q_{n-1}(s)$ in (14) by $s - c$ until another constant remainder r_1 is obtained, the value of the first derivative of $N(s)$ at $s = c$ will be found.

$$\frac{Q_{n-1}(s)}{s-c} = Q_{n-2}(s) + \frac{r_1}{s-c},$$

$$\frac{dN(s)}{ds}\Bigg]_{s=c} = r_1 .$$

Repeating this process, the value of the ith derivative of $N(s)$ at $s = c$ divided by factorial i is found to be equal to the ith constant remainder r_i

$$\frac{1}{i!}\cdot\frac{d^i N(s)}{ds^i}\Bigg]_{s=c} = r_i .$$

The verification of these relations may be found in [11, Chap. 3].

Denoting the remainder, obtained from the denominator $D(s)$ through the procedure indicated above, by r_i' , the expression for $Z(s + c)$ may be written as

$$Z(s + c) = K\frac{s^n + r_{n-1}s^{n-1} + \cdots + r_1 s + r_0}{s^m + r'_{m-1}s^{m-1} + \cdots + r'_1 s + r'_0}. \tag{15}$$

Therefore, for any given $Z(s)$, a c may be determined which produces a class k function $Z(s + c)$ by applying the tests for a class k function [1, Theorems 4 and 6] to expression (15).

If $Z(s)$ has only one finite zero, an application of [1, Theorem 5] yields values for c which insure that $Z(s + c)$ is class k. In this case, $Z(s + c)$ has the form

$$Z(s + c) = K\frac{s + c + a_0}{s^m + r'_{m-1}s^{m-1} + \cdots + r'_1 s + r'_0} = K\frac{s + c + a_0}{D(s + c)}, \tag{16}$$

where $r'_{m-1} = mc + b_{m-1}$, $r'_1 = dD(s)/ds$ at $s = c$, and $r'_0 = D(c)$. For m even, $Z(s + c)$ will be class $m - 1$ for all values of c greater than or equal to $(a_0 - b_{m-1})/(m - 1)$. For m odd, this theorem places two restrictions on c; it must be greater than or equal to $(a_0 - b_{m-1})/(m - 1)$ and greater than $(r'_0/r'_1 - a_0)$.

A simple graphical method for determining this constant c to insure a class k function $Z(s + c)$ is based on the fact that, if the phase angle of $Z(j\omega)$ equals an odd multiple of $\pi/2$ exactly $k - 1$ times for k odd or if this angle equals a multiple of π or zero exactly $k - 1$ times for k even and if $d\varphi/d\omega < 0$ at $\omega = 0$, then $Z(s)$ is class k [1, p. 282]. The procedure is quite simple if $N(s)$ and $D(s)$ are factored

$$Z(s) = K \frac{(s - \mu_1)(s - \mu_2) \cdots (s - \mu_n)}{(s - \rho_1)(s - \rho_2) \cdots (s - \rho_m)}. \tag{17}$$

Following the convention that the phase angle of any factor in (17) remains within the interval $(3\pi/2, -\pi/2)$, the phase angle of $(s - \mu_i)$ will be denoted by ψ_i and the phase angle of $(s - \rho_i)$ will be denoted by θ_i. The phase angle φ for $Z(s + c)$ is then given by

$$\varphi = \sum_{i=1}^{n} \psi_i - \sum_{i=1}^{m} \theta_i . \tag{18}$$

Now the phase angle of the factor $(s + c - \mu_i)$ is the angle of the vector from the point μ_i to the point $s + c$. Consequently, the variations of φ along a series of lines parallel to the imaginary axis in the s plane may be obtained graphically by applying (18). For some sufficiently large value of c, the aforementioned sufficiency criterion on the values of φ for $s = j\omega$ will be fulfilled and, for this value of c, $Z(s + c)$ will be class k.

Bounds on the impulse response of any transfer function. Now that any transfer function $Z(s)$ may be related to a class k function $Z(s + c)$ by an appropriate choice of c, the restrictions on the impulse responses of class k functions may be extended to the impulse response of $Z(s)$.

Theorem 7. The unit impulse response $W(t)$ corresponding to any transfer function $Z(s)$ is bounded by

$$| W(t) | \leq \frac{Ke^{ct}t^{k-1}}{(k - 1)!}, \tag{19}$$

where the constant c is such that $Z(s + c)$ is a class k function.

Proof. Consider the inverse Laplace transformation between $Z(s)$ and $W(t)$ where the path of integration is along the line $s = c + j\omega$

$$W(t) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} Z(s)e^{st}\, ds. \tag{20}$$

Making the change of variable $s = v + c$, (20) becomes

$$W(t) = e^{ct}W_c(t), \tag{21}$$

where

$$W_c(t) = \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} Z(v + c)e^{vt}\, dv. \tag{22}$$

Since $Z(v + c)$ is a class k function of v, Theorem 3 may be invoked to complete the proof. It should be noted that c need not be a positive number if $Z(s)$ is class k.

An application of [1, Theorem 10] to the quantity $W(t)/e^{ct}$ yields still another restriction on the unit impulse response of any transfer function.

Theorem 8. *Let* $| W(t)e^{-ct} |$ *be less than or equal to the positive constant* M *for* $t \geq \tau$. *For any transfer function* $Z(s)$, *let* c *be such that the imaginary part of* $Z_{k-2}(c + j\omega)$ *is non-positive for* $\omega \geq 0$. *Then, for* $0 \leq y < 1$,

$$| W(y\tau) | \leq e^{cy\tau}\left[\frac{K(y\tau)^{k-2} \sin \pi y}{\pi(k-1)!} + \frac{2My^k \sin \pi y}{\pi} \sum_{p=1}^{\infty} \frac{1}{p^{k-1}(p^2 - y^2)}\right].$$

Appendix. As an illustration, the bounds on the unit impulse response of the transfer function,

$$Z(s) = \frac{s + \mu}{(s + \rho)^3},$$

will be obtained. The only restriction imposed upon the real numbers μ and ρ is that ρ is positive. The unit impulse response corresponding to $Z(s)$ is

$$W(t) = e^{-\rho t}\left[t + \frac{\mu - \rho}{2} t^2\right]. \tag{23}$$

Now, the lowest real number c which makes $Z(s + c)$ a class 2 function will be found. According to the definition, $Z(s + c)$ will be class 2 when $Z_1(s + c)$ is positive real. But,

$$Z_1(s) = \int_s^{\infty} \frac{s_0 + \mu}{(s_0 + \rho)^3} ds_0 = \frac{s + \frac{\mu + \rho}{2}}{(s + \rho)^2}.$$

Calculating the real part of $Z_1(s + c)$, it can be seen that this real part is non-negative only when c satisfies the following inequalities

$$c \geq -\frac{\mu + \rho}{2} \quad \text{for} \quad \mu \leq \rho,$$

$$c \geq \frac{\mu - 3\rho}{2} \quad \text{for} \quad \mu \geq \rho.$$

Thus, by Theorem 7

$$| W(t) | \leq t \exp\left(-\frac{\mu + \rho}{2} t\right) \quad \text{for} \quad \mu \leq \rho$$

and

$$| W(t) | \leq t \exp\left(\frac{\mu - 3\rho}{2} t\right) \quad \text{for} \quad \mu \geq \rho.$$

By comparison with expression (23), it is easy to show that these inequalities are indeed true.

References

1. A. H. Zemanian, *On transfer functions and transients*, Quart. Appl. Math. **16**, 273–294 (1958)
2. M. F. Gardner and J. L. Barnes, *Transients in linear systems*, vol. 1; John Wiley and Sons, New York, 1942
3. D. F. Tuttle, Jr., *Network synthesis*, vol. 1, John Wiley and Sons, New York, 1958
4. M. Mathias, *Uber positive Fourier-Integrale*, Math. Z. **16**, 103–125 (1923)
5. M. Ky Fan, *Les fonctions definies—positives et les fonctions completement monotones*, Mem. Sci. Math., Acad. Sci. Paris, Fascicule 64 (1950)
6. H. Cramer, *On the representation of a function by certain Fourier integrals*, Trans. Am. Math. Soc. **46**, 191–201 (1939)
7. S. Bochner, *Monotone Funktionen, Stieltjessche Integrale und harmonische Analyse*, Math. Ann. **108**, 378–410 (1933)
8. K. Yosida, *On the representation of functions by Fourier integrals*, Proc. Imperial Acad., Tokyo, **20**, 655–660 (1944)
9. A. H. Zemanian, *Bounds existing on the time and frequency responses of various types of networks*, Proc. IRE **42**, 835–839 (1954)
10. A. H. Zemanian and P. E. Fleischer, *On the transient responses of ladder networks*, Trans. IRE, PGCT **CT-5**, 197–201 (1958)
11. L. Weisner, *Introduction to the theory of equations*, The Macmillan Co., New York, 1938

BOOK REVIEWS

Lineare algebra. By Werner Graeub. Springer-Verlag, Berlin, Gottingen, Heidelberg, 1958. x + 219 pp. $9.35.

A remarkably complete text on the subject of the title. By linear algebra is meant the theory of finite dimensional vector spaces and the linear and multilinear functions on them. This theory is conveniently subdivided into three parts, the general theory which applies to vector spaces over an arbitrary field, the theory of real vector spaces, and the theory of complex vector spaces. The principal results in all three branches are presented here. The book is divided into eleven chapters. The first two on "Lineare Räume" and "Lineare Abbildungen und Gleichungssysteme" are prerequisites for all the chapters which follow. Most of these latter, however, can be read independently of each other. Chapter III "Determinanten" is needed in order to read Chapter IV "Orientierte Lineare Räume" which deals with real vector spaces and is perhaps the most unusual part of the book. The main result of the chapter states that the basis x_ν can be deformed continuously into the basis y_ν if, and only if, the determinant of the mapping $x_\nu \to y_\nu$ is positive. Chapter V "Multilineare Algebra" is the longest in the book and treats the algebraic theory of tensors in an invarient (coordinate free) manner à la Bourbaki. The next four chapters are concerned with real vector spaces. Chapter VI "Der Euklidische Raum" introduces the scalar product. Chapter VII "Lineare Abbildung Euklidischer Räume" is concerned with Eigen values and the spectral theorem for self adjoint transformations. Chapter VIII "Symmetrische Bilinearfunktionen" treats quadratic forms and the inertia index. Chapter IX "Flächen zweiter Ordnung" presents both the affine and Euclidean classification of quadric surfaces. Chapter X "Unitäre Räume" deals with the spectral theorem for Hermitian and normal transformations. The final chapter XI "Invariante Unterräume" returns to the general theory giving an invarient treatment of canonical forms of transformation and the Cayley-Hamilton Theorem.

The book is suitable for a text, probably on the graduate level, and has many exercises in each chapter. I suspect it will be even more useful as a reference book in the subject for practicing mathematicians who are not specialists in this particular field.

DAVID GALE

Readings in linear programming. By S. Vajda. John Wiley & Sons, New York, 1958. vii + 99 pp. $3.00.

In this introduction to linear programming typical applications are illustrated by numerical examples which suggest the appropriate concepts and techniques. Only very elementary algebra is required for the understanding of the book, and the exposition is clear and easy to follow. The contents of the book are indicated by the following (incomplete) list of chapter headings: Transportation Problem, Caterer Problem, Production Scheduling, Transhipment, Bid Evaluation, Flow through a Network, Ship Scheduling, Personnel Assignment, Routing Aircraft, Investment, The Simplex Tableau, Nutrition Problem, Airlift, Blending of Aviation Gasolines, Smooth Patterns of Production, Duality, Selection of Products, Train Loss Reduction, Attendant's Rota, Warehousing, Games, Bibliography.

W. PRAGER

Calculus of variations and its applications. L. M. Graves (Editor). McGraw-Hill Book Company, Inc., New York, 1958. v + 153 pp. $7.50.

This important book contains the papers which were presented at the Eighth Symposium in Applied Mathematics sponsored by the American Mathematical Society and the Office of Ordnance Research, and held in April, 1956. The papers are noteworthy for their clarity of exposition and avoidance of an excessively cramped style, so that they should be of interest to physicists, engineers and operations analysts, as well as to mathematicians. They cover a wide range of topics in the calculus of variations and its applications to elasticity, plasticity, electromagnetic theory, mathematical economics and other fields, as the following list of authors and titles shows:

(*Continued on p. 262*)

HEAT CONDUCTION FROM A CYLINDRICAL SOURCE WITH INCREASING RADIUS*

BY

H. R. BAILEY

The Ohio Oil Co., Littleton, Colorado

Introduction. The theory of heat flow due to conduction from a moving heat source is of interest in a number of applications; for example, welding [8]**, heat exchangers [2, 3] and progressive freezing of a liquid [2, 3]. In this paper we consider heat conduction in an infinite homogeneous medium from the surface of a cylinder of finite length whose radius is increasing with time. This problem arises in connection with secondary oil recovery by an underground combustion process, e.g., [4, 7]. The restriction of the cylindrical source to a finite length corresponds to considering an oil reservoir of finite thickness and including vertical heat losses to media bounding the reservoir.

The differential equation describing this problem is written and the Greens function method is applied to obtain a solution in the form of an integral. A limiting value of this integral is then obtained for the case of the source moving at a constant velocity with no vertical losses. The problem is to evaluate a limit of the form $\lim_{t\to\infty} \int_0^t f(t, \tau)d\tau$ when $f(t, \tau)$ has an asymptotic representation, for $(t - \tau)/\tau$ sufficiently large, which can be integrated explicitly. For the integrand considered in Sec. 2, it is shown that the integral can be divided into two parts, namely

$$\int_0^t f(t, \tau)\, d\tau = \int_0^{t/N} f(t, \tau)\, d\tau + \int_{t/N}^t f(t, \tau)\, d\tau,$$

where the last integral goes to zero as $t \to \infty$ and the integrand in the range $[0, t/N]$ can be replaced by its asymptotic expression and evaluated in terms of N for $t \to \infty$. Finally, the desired limit is obtained by passing to the limit as $N \to \infty$.

In Sec. 3 an explicit evaluation of the integral solution is obtained for the case of the radius of the cylindrical source increasing at a variable velocity, namely $r_F' = V/r_F$. This result is obtained by showing that the solution in a transformed coordinate system is independent of time and thus the partial differential equation in the new coordinate system reduces to an ordinary differential equation which is solved explicitly.

1. A heat conduction problem. The partial differential equation in cylindrical coordinates for the temperature in a homogeneous conducting medium is given by

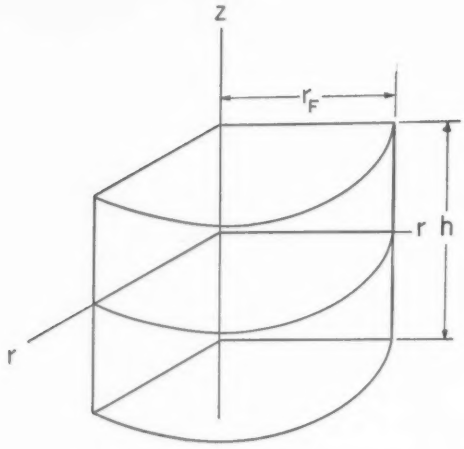
$$\frac{\partial^2 T}{\partial r^2} + \frac{1}{r}\frac{\partial T}{\partial r} + \frac{\partial^2 T}{\partial z^2} - a^2 \frac{\partial T}{\partial t} = -\frac{1}{k}\Phi(z, r, t), \tag{1}$$

where r is the radius, t is time, a^2 is reciprocal diffusivity, k is the conductivity, $\Phi(z, r, t)$ is the source function and T is the temperature which is independent of the angular position, θ. The source is assumed to be at the surface, $r = r_F(t)$, of a cylinder with a fixed axis at $r = 0$ and height h as shown in figure below.

---

*Received July 7, 1958.

**Numbers in brackets refer to the References at the end of the paper.



The intensity of the source is given by $L(t)$ in units of heat per unit of surface area per unit time. The source function, $\Phi(z, r, t)$, for this type of source may be expressed in terms of the Dirac delta function, $\delta(r - r_F)$, by the equation

$$\Phi(z, r, t) = L(t)B(z, h/2)\ \delta(r - r_F), \tag{2}$$

where $B(z, h/2) = 1$ for $|z| \leq h/2$, $= 0$ for $|z| > h/2$.

If we assume $T(z, r, 0) = 0$, then Eq. (1) has a solution of the form [see 1,7],

$$T(z, r, t) = (4\pi k)^{-1} \int_0^t dt_0 \int_{r_0=0}^{\infty} \int_{\theta_0=0}^{2\pi} \int_{z_0=-\infty}^{\infty} \Phi(z_0, r_0, t_0) G r_0\, dr_0\, d\theta_0\, dz_0, \tag{3}$$

where Φ is given by Eq. (2) and G is the Green's function corresponding to the left side of Eq. (1). G is given by the formula [see 1, 7],

$$G \equiv G(z, r, \theta, t \mid z_0, r_0, \theta_0, t_0)$$
$$= 2^{-1}a\pi^{-1/2}(t - t_0)^{-3/2} \exp\{-a^2[R^2 + (z - z_0)^2]/4(t - t_0)\},$$

where $R^2 = r^2 + r_0^2 - 2rr_0\cos(\theta - \theta_0)$.

After performing the indicated integrations in Eq. (3) with respect to z_0, θ_0 and r_0 and making the substitution $t - t_0 = \tau$ we obtain

$$T(z, r, t) = (4k)^{-1} \int_0^t d\tau\, \tau^{-1} L r_F \exp\left[\frac{-a^2(r^2 + r_F^2)}{4\tau}\right] I_0\left(\frac{a^2 r r_F}{2\tau}\right)$$
$$\cdot\left[\operatorname{erf}\frac{a(h/2 + z)}{2\sqrt{\tau}} + \operatorname{erf}\frac{a(h/2 - z)}{2\sqrt{\tau}}\right],$$

where L and r_F are evaluated at $t - \tau$, $I_0(z)$ is a modified Bessel function defined in Sec. 2.1 and $\operatorname{erf} x = 2\pi^{-1/2} \int_0^x \exp(-\alpha^2) d\alpha$.

A reasonable assumption in an underground combustion process is that $L = q\dot{r}_F$ where q is a constant. Two cases of this problem are of particular interest in underground combustion, namely: Case I = constant radial velocity, $r_F = Ut$, and Case II = frontal radius given by $r_F^2 = 2Vt$. The solutions for these two cases may be obtained from the

above equation; and for no vertical losses, i.e. $h \to \infty$, are given by Eqs. (4) and (5) respectively.

$$T(r, t) = (2k)^{-1} qU^2 \int_0^t d\tau (t - \tau)\tau^{-1} \exp\left\{\frac{-a^2[r^2 + U^2(t-\tau)^2]}{4\tau}\right\} I_0\left[\frac{a^2 rU(t-\tau)}{2\tau}\right] \quad (4)$$

$$T(r, t) = (2k)^{-1} qV \int_0^t d\tau\, \tau^{-1} \exp\left\{\frac{-a^2[r^2 + 2V(t-\tau)]}{4\tau}\right\} I_0\left\{\frac{a^2 r[2V(t-\tau)]^{1/2}}{2\tau}\right\}. \quad (5)$$

In Sec. 2 a steady state solution of (4) is obtained, that is, the $\lim_{t\to\infty} T(r, t)$, assuming that $s = r - Ut$ remains finite as $t \to \infty$. Thus a steady state solution for r positions near the source is determined; this has been called quasi-stationary state [see 6]. In Sec. 3 an explicit evaluation of the solution for Case II is obtained.

2. A steady state solution for the constant velocity case. 2.1. *Preliminary Considerations.* In this section we prove a lemma, state some known properties of $I_0(z)$, and discuss an explicit evaluation of an integral. These results will be needed in Sec. 2.2.

Lemma. Let $f(t)$ be a real function of the real variable t satisfying the conditions (α) $0 < f_2(N) \leq \lim_{t\to\infty} f(t) \leq f_1(N)$ for all finite N, and

$$(\beta) \lim_{N\to\infty} f_1(N) = \lim_{N\to\infty} f_2(N) = P, \quad \text{then} \quad \lim_{t\to\infty} f(t) = P.$$

Proof. By (β) there exists an N_0 such that $| f_1(N) - P | < \epsilon$ and $| f_2(N) - P | < \epsilon$ for all $N > N_0$. Combining this with (α) we have $P - \epsilon < \lim_{t\to\infty} f(t) < P + \epsilon$ and by choosing N_0 sufficiently large we can make ϵ small and the lemma is proved.

The modified Bessel function, $I_0(z)$, is defined by $I_0(z) = (2\pi)^{-1} \int_0^{2\pi} e^{z\cos\theta} d\theta$. $I_0(z)$, for real z, can be represented for large values of z by its asymptotic series $I_0(z) \sim (2\pi z)^{-1/2} e^z [1 + 0(1/z)]$, and thus

$$\lim_{z\to\infty} (2\pi z)^{1/2} e^{-z} I_0(z) = 1. \quad (6)$$

Since $e^{-z} I_0(z)$ is a continuous function of z and $\lim_{z\to\infty} e^{-z} I_0(z) = 0$, then $e^{-z} I_0(z)$ is bounded for $z \geq z_0$ for any z_0.

We define a function $g(z)$ by the equation

$$e^{-z} I_0(z) = (2\pi z)^{-1/2}[1 + g(z)], \qquad z \neq 0. \quad (7)$$

And if we define $g(0) = -1$, then $g(z)$ is a continuous function of z. Finally from (6) and (7) we have that $\lim_{z\to\infty} g(z) = 0$.

An explicit evaluation for the integral,

$$\psi(u, v, t) = \int_0^t x^{-1/2} \exp(-u^2/x) \exp(-v^2 x)\, dx, \qquad u \neq 0,$$

has been given by W. Horenstein [5]. The result is

$$\psi(u, v, t) = (2v)^{-1}\pi^{1/2}[-2 \sinh(2 | u | v) + e^{2|u|v} \operatorname{erf}(vt^{1/2} + | u | t^{-1/2}) + e^{-2|u|v} \operatorname{erf}(vt^{1/2} - | u | t^{-1/2})]. \quad (8)$$

The result of Horenstein does not include the absolute values signs on the u's and these must be added to make the formula correct for $u < 0$, thus we must have $\psi(u, v, t) = \psi(-u, v, t)$. In the application of this formula in Sec. 2.2, the case $u < 0$ corresponds to positions inside the cylinder, i.e. $s < 0$. Passing to the limit as $t \to \infty$ in (8) we obtain

$$\lim_{t\to\infty} \psi(u, v, t) = v^{-1}\pi^{1/2}e^{-2|u|v}. \tag{9}$$

2.2. *Evaluation of a limit.* Equation (4) may be written in the form

$$T = (2k)^{-1}qU^2I, \tag{10}$$

where

$$I = \int_0^t d\tau(t-\tau)\tau^{-1} \exp\{-a^2[r - U(t-\tau)]^2/4\tau\} \exp[-a^2rU(t-\tau)/2\tau] \cdot I_0[a^2rU(t-\tau)/2\tau]$$

and putting $s = r - Ut$, $U > 0$, we obtain

$$I = \int_0^t d\tau(t-\tau)\tau^{-1} \exp[-a^2(s+U\tau)^2/4\tau] \exp[-z(t,\tau)]I_0[z(t,\tau)], \tag{11}$$

where

$$z(t,\tau) = (2\tau)^{-1}a^2U(s+Ut)(t-\tau). \tag{12}$$

We shall temporarily assume $s \neq 0$ and in this case the integrand in (11) is bounded in $0 \leq \tau \leq t$, since for $\tau \neq 0$ the integrand is a product of continuous functions of τ and the integrand approaches zero as $\tau \to 0$.

I may be divided into two parts, $I = I_1 + I_2$, where $I_1 = \int_0^{t/N} f(t,\tau)d\tau$ and $I_2 = \int_{t/N}^t f(t,\tau)d\tau$ with $f(t,\tau)$ written for the integrand in (11), where N is chosen > 1. For $0 < t/N \leq \tau \leq t$, we have $0 \leq (t-\tau)/\tau \leq (t - t/N)/(t/N) = N - 1$. Then it follows from Eq. (12), with $U > 0$ and s finite that $z(t,\tau)$ is bounded below for $0 < t/N \leq \tau < t$ and $N > 1$, say $z(t,\tau) \geq z_0$ and thus, by Sec. 2.1, that $e^{-z(t,\tau)}I_0[z(t,\tau)]$ is bounded. Hence there exists an M such that $0 < e^{-z(t,\tau)}I_0[z(t,\tau)] \leq M$ for $0 < t/N \leq \tau \leq t$, $N > 1$. The left side of the above inequality is true since $I_0(z) \geq 1$ for all z. Using the above evaluations we can now evaluate I_2. I_2 can be written in the form

$$I_2 = \int_{t/N}^t d\tau\, \tau^{-1}(t-\tau) \exp(-a^2s^2/4\tau) \exp(-a^2sU/2) \cdot \exp(-a^2U^2\tau/4) \exp[-z(t,\tau)]I_0[z(t,\tau)]$$

and thus

$$0 \leq I_2 \leq (N-1)M \exp(-a^2sU/2) \int_{t/N}^t \exp(-a^2U^2\tau/4)\, d\tau$$
$$= (N-1)M \exp(-a^2sU/2)(-4)(aU)^{-2}[\exp(-a^2U^2t/4) - \exp(-a^2U^2t/4N)]$$

and thus for $t \to \infty$ and for all finite N we have

$$\lim_{t\to\infty} I_2(t, N) = 0. \tag{13}$$

If we replace $e^{-z(t,\tau)}I_0[z(t,\tau)]$ in I_1 by its equivalent given in Eq. (7) we obtain

$$I_1 = \int_0^{t/N} d\tau(t-\tau)\tau^{-1} \exp[-a^2(s+U\tau)^2/4\tau]\{1 + g[z(t,\tau)]\}[2\pi z(t,\tau)]^{-1/2}$$
$$= [a^2U\pi(s+Ut)]^{-1/2} \int_0^{t/N} d\tau(t-\tau)^{1/2}\tau^{-1/2} \cdot \exp[-a^2(s+U\tau)^2/4\tau]\{1 + g[z(t,\tau)]\}, \tag{14}$$

where $z(t, \tau)$ in the last factor of the integrand of the first equation of (14) has been replaced by its expression in Eq. (12). If we define

$$h(t, \tau) = (t - \tau)^{1/2}\tau^{-1/2} \exp [-a^2(s + U\tau)^2/4\tau], \qquad h(t, 0) = 0 \quad \text{and} \quad g[z(t, 0)] = 0,$$

then the integrand of the second equation of (14) can be written as the product of $h(t, \tau)$ and $1 + g[z(t, \tau)]$, where both factors are continuous functions of τ; and $h(t, \tau) \geq 0$ for $0 \leq \tau \leq t/N$ and for all $t \geq 0$. Hence we may apply the first mean value theorem for integrals to the integral in the second equation of (14) and obtain

$$I_1 = [a^2U\pi(s + Ut)]^{-1/2}\{1 + g[z(t, \xi)]\} \int_0^{t/N} d\tau(t - \tau)^{1/2}\tau^{-1/2} \exp [-a^2(s + U\tau)^2/4\tau],$$

where $0 \leq \xi \leq t/N$. For $0 \leq \tau \leq t/N$ we have $(t)^{1/2}(1 - 1/N)^{1/2} \leq (t - \tau)^{1/2} \leq (t)^{1/2}$ and thus

$$\begin{aligned} I_1 &\leq t^{1/2}[a^2U\pi(s + Ut)]^{-1/2}\{1 + g[z(t, \xi)]\}J(t, N) \\ I_1 &\geq t^{1/2}(1 - 1/N)^{1/2}[a^2U\pi(s + Ut)]^{-1/2}\{1 + g[z(t, \xi)]\}J(t, N), \end{aligned} \tag{15}$$

where

$$J(t, N) = \int_0^{t/N} d\tau\, \tau^{-1/2} \exp [-a^2(s + U\tau)^2/4\tau].$$

For $0 \leq \xi \leq t/N$, we have $(t - \xi)/\xi \geq (t - t/N)/(t/N) = N - 1$; then

$$z(t, \xi) = 2^{-1}a^2U(s + Ut)(t - \xi)/\xi \geq 2^{-1}a^2U(s + Ut)(N - 1),$$

and thus $\lim_{t\to\infty} z(t, \xi) = \infty$ and finally we have $\lim_{t\to\infty} g[z(t, \xi)] = \lim_{z\to\infty} g(z) = 0$. If we now pass to the limit as $t \to \infty$ in (15) we obtain

$$(1 - 1/N)^{1/2}(aU)^{-1}\pi^{-1/2} \lim_{t\to\infty} J(t, N) \leq \lim_{t\to\infty} I_1 \leq (aU)^{-1}\pi^{-1/2} \lim_{t\to\infty} J(t, N) \tag{16}$$

and evaluating $\lim_{t\to\infty} J(t, N)$, we have

$$\begin{aligned} \lim_{t\to\infty} J(t, N) &\equiv \lim_{t\to\infty} \int_0^{t/N} d\tau\, \tau^{-1/2} \exp [-a^2(s + U\tau)^2/4\tau] \\ &= \exp (-a^2sU/2) \int_0^{\infty} d\tau\, \tau^{-1/2} \exp (-a^2s^2/4\tau) \exp (-a^2U^2\tau/4) \end{aligned}$$

and from Eq. (9) we have

$$\lim_{t\to\infty} J(t, N) = \exp (-a^2sU/2)2(aU)^{-1}(\pi)^{1/2} \exp (-a^2 \mid s \mid U/2). \tag{17}$$

Combining (13), (16) and (17), we have

$$\begin{aligned} (1 - 1/N)^{1/2}2(aU)^{-2} \exp [-a^2U(s + \mid s \mid)/2] &\leq \lim_{t\to\infty} (I_1 + I_2) \\ &\leq 2(aU)^{-2} \exp [-a^2U(s + \mid s \mid)/2] \end{aligned}$$

and the above inequality is of the form discussed in the lemma of Sec. 2.1 and thus we may pass to the limit as $N \to \infty$ and obtain

$$\lim_{t\to\infty} I = 2(aU)^{-2} \exp [-a^2U(s + \mid s \mid)/2]. \tag{18}$$

In the above calculations we have assumed $s \neq 0$; if $s = 0$, we may replace the lower limit of the integral in Eq. (11) by ϵ, repeat the calculations and then pass to the limit as $\epsilon \to 0$; and we obtain (18) for the case $s = 0$.

If we combine Eqs. (8) and (18) we obtain for $t \to \infty$

$$\lim_{t\to\infty} T \equiv 2k^{-1}qU^{-2} \lim_{t\to\infty} \int_0^t (t - \tau)\tau^{-1} \cdot \exp\{-a^2[r^2 + U^2(t - \tau)^2]/4\tau\} I_0[a^2 rU(t - \tau)/2\tau]\, d\tau \qquad (19)$$

$$= qk^{-1}a^{-2} \exp[-a^2U(s + |s|)/2].$$

The above result is the same as the corresponding known result, [see 6], for a plane source moving with a constant velocity. This would be expected since the cylindrical source approaches a plane source for large radii.

3. An explicit solution for Case II. Substituting $x = r/r_F$ and $\tau = \lambda t$ in Eq. (5) gives

$$T(x, t) = (2k)^{-1}qv \int_0^1 \frac{d\lambda}{\lambda} \exp\left[\frac{-a^2V(x^2 + 1 - \lambda)}{2\lambda}\right] I_0\left[\frac{a^2Vx(1 - \lambda)^{1/2}}{\lambda}\right]. \qquad (20)$$

Thus T, when written as a function of x, is independent of t. The partial differential equation, Eq. (1), for the corresponding case may be written in the form

$$\frac{\partial^2 T}{\partial x^2} + \left(\frac{1}{x} + a^2Vx\right)\frac{\partial T}{\partial x} - 2a^2Vt\,\frac{\partial T}{\partial t} = 0, \qquad (21)$$

where we have made the substitution $x = r/r_F$. The term $\partial^2 T/\partial z^2 = 0$, since, for the case $h \to \infty$, T is independent of z. The source term is replaced by its equivalent condition, namely

$$\left.\frac{\partial T}{\partial x}\right|_{x=1+} - \left.\frac{\partial T}{\partial x}\right|_{x=1-} = -(k)^{-1}r_F L(t) = -qV/k, \qquad (22)$$

where

$$\left.\frac{\partial T}{\partial x}\right|_{x=1+}, \qquad \left.\frac{\partial T}{\partial x}\right|_{x=1-}$$

are the right and left hand derivatives, respectively, at $x = 1$.

Due to symmetry no temperature gradient exists at $r = 0$, also $T \to 0$ as $r \to \infty$. The corresponding boundary conditions on $T(x, t)$ are

$$\frac{\partial T}{\partial x}(0, t) = 0; \qquad \lim_{x\to\infty} T(x, t) = 0. \qquad (23)$$

Thus Eq. (20) is a solution of Eq. (21) subject to conditions (22) and (23). Since the solution (20) is independent of t, then $\partial T/\partial t = 0$ and Eq. (21) reduces to the ordinary differential equation

$$x\,\frac{d^2T}{dx^2} + (1 + a^2vx^2)\,\frac{dT}{dx} = 0. \qquad (24)$$

Equation (24) can be integrated giving the two solutions

$$T = \text{constant}$$

$$T = C_1 \int_{C_2}^{x} \frac{dz}{z} \exp\left[-a^2 V z^2/2\right].$$

To satisfy the boundary condition, $dT/dx = 0$ for $x = 0$, we must choose the solution $T = \text{constant}$ for $0 \leq x \leq 1$; thus the boundary condition corresponding to (22) becomes

$$\left.\frac{dT}{dx}\right|_{x=1+} = \frac{-qV}{k}$$

and the solution in the range $x \geq 1$ is given by

$$\begin{aligned} T &= -qV(k)^{-1} \exp\left(a^2 V/2\right) \int_{\infty}^{x} d(z)(z)^{-1} \exp\left(-a^2 V z^2/2\right) \\ &= -(2k)^{-1} qV \exp\left(a^2 V/2\right) Ei(-a^2 V x^2/2), \qquad x \geq 1 \end{aligned} \tag{25}$$

where C_1 and C_2 have been chosen so that (25) will satisfy the boundary conditions at $x = 1$ and $x \to \infty$. Finally, the value of the constant temperature in the range $0 \leq x \leq 1$ is obtained by requiring that the solution be continuous at $x = 1$ thus

$$T = -(2k)^{-1} qV \exp\left(a^2 V/2\right) Ei(-a^2 V/2), \qquad 0 \leq x \leq 1 \tag{26}$$

where Ei is the exponential integral, $-Ei(-x) = \int_x^{\infty} e^{-z} z^{-1} dz$. Returning to the r coordinate, by putting $r = x r_F(t)$ in the solution (25) and (26), gives the desired explicit solution for Case II.

Since this solution is unique, we may equate the solution (20) to the solution (25), (26) and obtain the following interesting evaluation, with $a^2 V/2 = R$,

$$\int_0^1 \frac{d\lambda}{\lambda} \exp\left[\frac{-R(x^2+1)}{\lambda}\right] \cdot I_0\left[\frac{2Rx(1-\lambda)^{\frac{1}{2}}}{\lambda}\right] = \begin{cases} -Ei(-R), & |\, x \,| \leq 1 \\ -Ei(-Rx^2), & |\, x \,| \geq 1. \end{cases} \tag{27}$$

References

1. H. S. Carslaw and J. C. Jaeger, *Operational methods in applied mathematics*, Oxford Press, 1941
2. J. Crank, *The mathematics of diffusion*, Oxford Press, 1956
3. P. V. Dankwerts, *Unsteady state diffusion or heat conduction with a moving boundary*, Trans. Faraday Soc. **47**, 701–12 (1950)
4. Don V. Hester and D. E. Menzie, *Development of subsurface combusion drive*, Petroleum Eng. **26**, B-82, B-87-92 (Nov. 1954)
5. W. Horenstein, *On certain integrals in the theory of heat conductions*, Quart. Appl. Math. **3**, 183–84 (1945)
6. Max Jacob, *Heat transfer*, vol. I, Wiley & Sons, New York, 1949, p. 343–52
7. P. M. Morse and H. Feshbach, *Methods of theoretical physics*, McGraw-Hill, 1953, New York, p. 857–95
8. D. Rosenthal, *The theory of moving sources of heat and its application to metal treatments*, Trans. Am. Soc. Mech. Engrs. **68**, 849–65 (1946)

BOOK REVIEWS

(*Continued from p. 254*)

1. E. Reissner, "On Variational Principles in Elasticity"
1. D. C. Drucker, "Variational Principles in the Mathematical Theory of Plasticity"
3. P. G. Hodge, Jr., "Discussion of D. C. Drucker's Paper 'Variational Principles in the Mathematical Theory of Plasticity' "
4. J. B. Keller, "A Geometric Theory of Diffraction"
5. J. B. Diaz, "Upper and Lower Bounds for Eigenvalues"
6. J. L. Synge, "Stationary Principles for Forced Vibrations in Elasticity and Electromagnetism"
7. H. F. Weinberger, "A Variational Computation Method for Forced-Vibration Problems"
8. M. M. Schiffer, "Applications of Variational Methods in the Theory of Conformal Mapping"
9. R. Bellman, "Dynamic Programming and Its Application to Variational Problems in Mathematical Economics"
10. S. Chandrasekhar, "Variational Methods in Hydrodynamics"
11. E. H. Rothe, "Some Applications of Functional Analysis to the Calculus of Variations"

Unfortunately, only a one-page index is provided.

The personal tastes of the reviewer governed the choice of three of the longer papers for further discussion.

Keller introduces a generalized theory of geometrical optics designed to include the effects of diffraction. This is accomplished through the introduction of new rays, diffracted rays, which account for the appearance of light in shadow regions and which alter the light found in illuminated regions. These diffracted rays arise when a ray impinges on an edge or a vertex, or grazes an interface. First the diffracted rays are specified explicitly by describing the rays which arise under various circumstances. Then the rays are characterized by an extension of Fermat's principle, with the equivalence of the two descriptions following from standard variational considerations. Following this diffracted wavefronts, eiconal functions, and imaginary rays (which account for the light found in the region on that side of a caustic through which no ordinary rays pass) are introduced. Finally, to achieve a quantitative theory, amplitude and phase functions are associated with the rays. The author states, justifiably, "if this ray theory had been available at the time of the controversy between ray and wave theory, it might have forestalled the acceptance of the latter." The paper concludes with an interesting section relating this theory to earlier works in diffraction theory, along with an extensive bibliography.

The determination of approximate values for the characteristic numbers of self-adjoint differential operators is of great practical importance. The Rayleigh-Ritz method furnishes upper bounds and the method of Weinstein lower bounds. Using the equation for the vibrating clamped plate as an example, Diaz states Courant's minimax characterization of the eigenvalues and notes that by contracting the class of admissible functions upper bounds are obtained, and by enlarging it, lower bounds, which leads to a unified view of the basic ideas underlying these two methods of approximation. Then each is discussed in detail, along with Aronszajn's extensions. The paper is rounded out with some remarks concerning estimates of the errors and relations between the eigenvalues for the vibrating plate and vibrating membrane problems. Though no numerical examples are given, many references to the literature are provided which still further enhance the value of the paper.

The purpose of Bellman's paper is the discussion of a variety of optimization problems arising in mathematical economics. The approach is via the functional equation technique of dynamic programming, a field in which the author has pioneered and which has been extensively explored in recent years. Interest centers on the analytical and computational determination of optimal policies for use in multistage decision processes. The key to the solution of these problems is provided by the principle of optimality: an optimal policy has the property that whatever the initial state of the system is, and whatever the state is that results from the initial decision, the remaining decisions must constitute an optimal policy with regard to the state resulting from the initial decision. In turn, this is but a special case of the more general principle of invariant imbedding (*Proc. Nat. Acad. Sci.*, v. 42 (1956), p. 629), which can be used in treating a number of problems in mathematical physics. Following a brief outline of dynamic programming and the relationships between continuous decision processes and the calculus of variations, several multi-stage allocation processes and smoothing processes (of both deterministic

(*Continued on p. 270*)

IMPROVING THE CONVERGENCE IN AN EXPANSION OF SPHEROIDAL WAVE FUNCTIONS*

BY

J. MEIXNER (*Technische Hochschule Aachen*)

AND

C. P. WELLS (*Michigan State University*)

1. Introduction. Spheroidal configurations have proved to be of considerable value in various types of diffraction and radiation problems. Of these we mention only one example, the prolate spheroidal radiating antenna. This has been discussed recently by Myers [1] who studied the radiation patterns and by Wells [2] who studied the near field of the antenna. Other examples and references to the literature can be found in the books of Meixner and Schafke [3] and of Flammer [4].

The advantages of spheroidal models both in a mathematical and in a physical sense are well known and will not be discussed here. The disadvantages are the complexity of the spheroidal functions and the lack of sufficient numerical values of the functions. Further, the expansions in spheroidal functions which represent the near field, e.g. the current on the antenna, converge slowly, a fact that is also true of expansions for other geometric models. It is this slow convergence which will concern us in this note and we shall show that the convergence can be improved in the sense that fewer terms of the expansion are needed to obtain the desired accuracy. The method is not new; it has been used, e.g. by Meixner and Kloepfer [5] in the problem of the ring shaped antenna. However in the case of the spheroidal functions some device such as this is particularly useful and we present it here with the idea that it may be of use to others.

By improving convergence we mean the following: Let $\sum a_n\psi_n$ be an expansion of eigen functions ψ_n and consider the asymptotic behaviour of $a_n\psi_n$ for large n. Suppose that to orders of $1/n$, $a_n\psi_n \approx b_n\varphi_n$. Then in the expansion

$$\sum (a_n\psi_n - b_n\varphi_n) \tag{1}$$

the terms of order $1/n$ are absent and this sum can be expected to converge faster than the original expansion. If in addition it is possible to write the sum $\sum b_n\varphi_n$ in closed form, then

$$\sum a_n\psi_n = \sum (a_n\psi_n - b_n\varphi_n) + f, \tag{2}$$

where f is the sum $\sum b_n\varphi_n$. The results of this procedure, based on the prolate spheroidal antenna, will be given in what follows. Obviously the method can be applied to other types of eigen function expansions.

2. Notation and expansions. A variety of notations have been used in spheroidal functions most of which have been summarized in [4]. For theoretical work the notation used in [3] is preferable while for numerical work other notations have some advantage. We shall use the notation of [3] where also the relevant information concerning spheroidal functions can be found.

*Received September 12, 1958.

The spheroidal coordinates are ξ, η, φ where ξ is the radial variable and η and φ are angular variables. The ranges of the variables are $\xi > 1$, $| \eta | \leq 1$, $0 \leq \varphi \leq 2\pi$. The angular functions for the symmetrically driven antenna are $ps_n^1(\eta; \gamma^2)$, where $\gamma = 2\pi a/\lambda$, a is the semi-focal length of the spheroid $\xi = $ const. and λ is the wave length. The radial functions are $S_n^{1(4)}(\xi; \gamma)$ and are functions of the third kind in that they provide the proper wave function behaviour for large ξ. The φ-component of the magnetic field is given by

$$H_\varphi = \sum_{n=1}^{\infty} b_n S_n^{1(4)}(\xi; \gamma) ps_n^1(\eta; \gamma^2), \tag{3}$$

where b_n is given by

$$2[(\xi_0^2 - 1)^{1/2} S_n^{1(4)}(\xi_0; \gamma)]' b_n/(2n + 1) = i\omega\epsilon a \int_{\eta_1}^{\eta_2} E_\eta(\xi_0; \eta)(\xi_0^2 - \eta^2)^{1/2} ps_n^{-1}(\eta; \gamma^2)\, d\eta$$

and where $\xi_0 = $ const. represents the antenna, E_η is the tangential component of the electric field and the prime indicates differentiation with respect to ξ_0.

Equation (3) is our basic expansion. An indication of how slowly it converges for $\xi = \xi_0$ is given in Table 4. Since the spheroidal functions are difficult at best to compute and existing tables so far are limited in range, some method which cuts down this computing and speeds up convergence is desirable.

3. Comparison series. Ignoring the physical constants and the integral over the gap in (3), the problem is the convergence of

$$V = \sum_{n=1}^{\infty} ps_n^{-1}(\eta'; \gamma^2) ps_n^1(\eta; \gamma^2)(n + 1/2) V_n(\xi), \tag{4}$$

where

$$V_n(\xi) = S_n^{1(4)}(\xi; \gamma)/[(\xi^2 - 1)^{1/2} S_n^{1(4)}(\xi; \gamma)]'.$$

We now consider some comparison series formed by replacing the terms of (4) by their asymptotic expressions for large n. The details of obtaining these expressions are outlined in the Appendix. According to Lemma I in the Appendix we can choose as a comparison series

$$W = \sum_{n=1}^{\infty} P_n^{-1}(\eta') P_n^1(\eta)(n + 1/2) W_n(\xi) \tag{5}$$

with

$$W_n(\xi) = (\xi^2 - 1)^{1/2} Q_n'(\xi)/n(n + 1) Q_n(\xi),$$

and where P_n^1, P_n^{-1}, Q_n are Legendre functions. Numerically the comparison of V_n with W_n depends on γ and the smaller the γ the better the comparison. Table 1 shows this comparison for $\gamma = 1$ and $\gamma = 2$ and for two different values of ξ. In this table only the the real part of V_n is given; the imaginary part falls off rapidly as n increases.

One concludes from Table 1 that (5) is, in fact, a very good comparison series if only γ is not too large. The difficulty in using it lies in the fact that W cannot be summed in closed form. But even so there is some practical value in replacing the terms of (4) with those of (5) with some accuracy after, say, $n = 6$, especially in view of the limited tables of spheroidal functions. The functions $Q_n(\xi)$ at least for values of ξ near 1 as well as the functions $P_n^1(\eta)$ can easily be computed.

TABLE 1.

$\xi = 1.00001$

	$\gamma = 1$		$\gamma = 2$	
n	$\mathrm{Re}(V_n)$	W_n	$\mathrm{Re}(V_n)$	W_n
1	−37.16	−21.923	29.396	−21.923
2	−8.922	−8.098	−13.199	−8.098
3	−4.548	−4.362	−5.235	−4.362
4	−2.845	−2.774	−3.065	−2.774
5	−1.979	−1.946	−2.058	−1.946
6	−1.491	−1.477	−1.521	−1.477
7		−1.163	−1.141	−1.163

$\xi = 1.001$

	$\gamma = 1$		$\gamma = 2$	
n	$\mathrm{Re}(V_n)$	W_n	$\mathrm{Re}(V_n)$	W_n
1	−6.763	−3.962	7.322	−3.962
2	−1.744	−1.586	−2.593	−1.586
3	−.947	−.911	−1.080	−.911
4	−.626	−.613	−.669	−.613
5	−.456	−.450	−.476	−.450
6	−.353	−.351	−.364	−.351
7			−.292	−.286

From the comparison series W, with the help of Lemmas II and III in the appendix, we obtain as an approximation to (5)

$$X = \sum_{n=1}^{\infty} P_n^{-1}(\eta')P_n^1(\eta)(n + 1/2)X_n(\xi)/n(n + 1), \tag{6}$$

where

$$X_n(\xi) = -\tfrac{1}{2} \coth u - (n + 1/2)[K_1[(n + 1/2)u]/K_0[(n + 1/2)u] - 1 - 1/(2n + 1)u - n(n + 1)/(n + 1/2) - 1/(4n + 2)]$$

with $\xi = \cosh u$. Let us put $X = X_1 + X_2 + X_3 + X_4$ and

$$X_i = \sum_{n=1}^{\infty} P_n^{-1}(\eta')P_n^1(\eta)X_{ni}(\xi), \qquad i = 1, 2, 3, 4$$

where

$$X_{n1} = -\tfrac{1}{2}[(n + 1/2)/n(n + 1)] \coth u,$$

$$X_{n2} = -1,$$

$$X_{n3} = -1/4n(n + 1),$$

$$X_{n4} = -[(n + 1/2)^2/n(n + 1)][K_1(u_n)/K_0(u_n) - 1 - 1/(2n + 1)u]$$

and $u_n = (n + 1/2)u$. The coefficients are of the order $X_{n1} = 0\ (n^{-1})$, $X_{n3} = 0\ (n^{-2})$, $X_{n4} = 0\ (n^{-3})$. Thus it would appear that it would be sufficient to take only X_2 as a comparison series in order to improve convergence. However this improvement becomes effective only after the coefficients of (6) with X_{n2} in place of X_n approximate those of (5) closely enough. And this is the case only for rather large values of n if ξ is very close to 1. In order to approximate the terms for smaller n also it is useful to take $X_1 + X_2 + X_3 + X_4$ as the comparison series.

As Table 2 shows X_{n3} is never very essential and can be omitted. For $\xi = 1.081$, the main contribution is given by X_{n2} although the additional consideration of X_{n1} and X_{n4} is useful. For ξ very close to 1, X_{n1} and X_{n4} are of much greater importance than X_{n2}.

TABLE 2.

	$\xi = 1.081$				$\xi = 1.02$			
	$u = 0.4$, coth $u = 2.632$				$u = 0.2$, coth $u = 5.067$			
n	X_{n1}	X_{n2}	X_{n3}	X_{n4}	X_{n1}	X_{n2}	X_{n3}	X_{n4}
1	−0.988	−1	−0.125	0.181	−1.900	−1	−0.125	0.495
2	−0.548	−1	−0.042	0.077	−1.056	−1	−0.042	0.220
3	−0.385	−1	−0.021	0.041	− .739	−1	−0.021	0.122
4	−0.296	−1	−0.013	0.025	−.570	−1	−0.013	0.081
5	−0.241	−1	−0.008	0.017	−.465	−1	−0.008	0.061
6	−0.204	−1	−0.006	0.012	−.392	−1	−0.006	0.050

	$\xi = 1.001$				$\xi = 1.00001$			
	$u = 0.0447$, coth $u = 22.38$				$u = 0.0047$, coth $u = 223.6$			
n	X_{n1}	X_{n2}	X_{n3}	X_{n4}	X_{n1}	X_{n2}	X_{n3}	X_{n4}
1	−8.40	−1	−0.125	3.64	−84.0	−1	−0.125	46.5
2	−4.67	−1	−0.042	1.76	−46.7	−1	−0.042	26.3
3	−3.27	−1	−0.021	1.11	−32.7	−1	−0.021	18.0
4	−2.52	−1	−0.013	0.76	−25.2	−1	−0.013	13.8
5	−2.05	−1	−0.008	0.58	−20.5	−1	−0.008	10.8
6	−1.73	−1	−0.006	0.47	−17.3	−1	−0.006	9.0

In the next section we shall show that X_1 and X_2 can be written in closed form. The series X_3 can be neglected in comparison with X_2. The series X_4 is more difficult to handle since it cannot be summed in closed form. However as Lemma IV shows, good approximations exist for both large and small values of ξ. Some values of K_1/K_0 and approximations for large and small argument are given in Table 3.

TABLE 3.

u_n	$K_1(u_n)/K_0(u_n)$	$u_n^{-1}[\gamma + \ln u_n/2]^{-1}$	$1 + 1/2u_n$	$1 + \frac{1}{2u_n} - K_1/K_0$
0.02	12.4	−12.4	26	13.6
0.04	7.5	−7.5	13.5	6.0
0.10	4.05	−4.12	6.0	1.95
0.20	2.73	−3.0	3.50	0.77
0.40	1.97		2.25	0.28
1.00	1.43		1.50	0.07
2.00	1.23		1.25	0.02

Returning to the original expansion (4), it can be seen from the coefficients $V_n(\xi)$ that the convergence improves as ξ approaches 1. However it does not follow that the comparison series give the best results for ξ in the neighborhood of 1. If we study these comparison series for various values of ξ we find that X_4 can be neglected in comparison with X_1 and X_2 if $(n + 1/2)u > 0.5$. This holds for all n if $\xi > 1.54$. For smaller values of ξ, X_4 plays an essential role. Thus for $\xi = 1.00001$, X_1 and X_4 give the main contribution for $n < 0.5/u$ or $n < 100$, approximately. This can be seen from studying the coefficients X_{n1} and X_{n4} in comparison with X_{n2}. Hence if we are interested in values of $\xi < 1.54$, the series X_4 is important and since this series cannot be summed we must compute or approximate the sum numerically. This, however, is not difficult in view of the simple analytic form of the coefficients X_{n4} and their approximations as given in Table 3.

4. Closed form expressions for X_1 and X_2. We consider first X_1 and notice that this series is the Green's function for Legendre's differential equation

$$(1 - \eta^2)y'' - 2\eta y' + [n(n + 1) - m^2/(1 - \eta^2)]y = 0$$

with $n = 0$ and $m = 1$. If we label this Green's function as $G(\eta, \eta')$ then

$$\begin{aligned} G(\eta, \eta') &= \tfrac{1}{2}[(1 + \eta)(1 - \eta')/(1 - \eta)(1 + \eta')]^{1/2}, \qquad \eta \leq \eta' \\ &= \tfrac{1}{2}[(1 - \eta)(1 + \eta')/(1 + \eta)(1 - \eta')]^{1/2}, \qquad \eta \geq \eta' \end{aligned}$$

when expressed in terms of the solutions $[(1 + \eta)/(1 - \eta)]^{\pm 1/2}$ of the differential equation.

In order to sum X_2 we make use of a result of Watson [6] on summation of the Gegenbauer functions. This result, specialized for our purposes, states that

$$\sin\theta \sin\varphi \int_0^\pi \frac{\sin^2 w\, dw}{(1 - 2t\cos\Omega + t^2)^{3/2}} = -\pi \sum_{n=0}^{\infty} t^n P_{n+1}^1(\cos\theta) P_{n+1}^{-1}(\cos\varphi) \qquad (7)$$

where $\cos\Omega = \cos\theta\cos\varphi + \sin\theta\sin\varphi\cos w$. If we make the substitution $w = \pi - 2\psi$, the integral on the left becomes

$$2\int_0^{\pi/2} \frac{\sin^2 2\psi\, d\psi}{(1 - 2t\cos\Omega + t^2)^{3/2}}.$$

For $t = 1$ this becomes

$$\frac{2}{a^{3/2}} \int_0^{\pi/2} \frac{\sin^2 2\psi\, d\psi}{(1 - k^2\sin^2\psi)^{3/2}} = \frac{8}{a^{3/2}k^4}[(2 - k^2)K - 2E],$$

where $a = 2 - 2\cos(\theta + \varphi)$, $k^2 = 2\sin\theta\sin\varphi/[1 - \cos(\theta + \varphi)]$ and K and E are the complete elliptic integrals. The integrals fail to exist when $\theta = \varphi$.

Now the expansion on the right of (7) converges for $|t| < 1$ and since $P_n^1(\cos\theta)P_n^{-1}(\cos\varphi)$ is $0(1/n)$ for large n and for θ and φ between 0 and π, the expansion converges also for $t = 1$. Then by Abel's theorem [7] the equality in (7) holds for $t = 1$ except for $\theta = \varphi$. Hence we have

$$X_2 = -\sum_{n=1}^{\infty} P_n^1(\cos\theta)P_n^{-1}(\cos\varphi) = (1/\pi k^2)\csc[(\theta + \varphi)/2][(2 - k^2)K - 2E]$$

with $\eta = \cos\theta$, $\eta' = \cos\varphi$.

5. Some numerical results on convergence. In Table 4 we give some numerical results which offer some indication of how well the convergence can be improved. The first six terms of the series for V and for X as well as for their difference are given for $\gamma = 2$ and for $\xi = 1.001$ and $\xi = 1.00001$; for $\gamma = 1$ only the first five terms are given with $\xi = 1.02$. Such a table is limited by the amount of information available on numerical values of the spheroidal functions and the above values of ξ were chosen since for these we have sufficient information to make a reasonable comparison of V and X. No attempt has been made here to tabulate X_1 and X_2 from their closed form expressions although such computations can readily be done.

TABLE 4.

$\gamma = 2$						$\gamma = 1$		
$\xi = 1.00001$			$\xi = 1.001$			$\xi = 1.02$		
V	X	V-X	V	X	V-X	V	X	V-X
17.05	−16.6	33.65	4.25	−2.54	6.79	−1.695	−1.09	−.605
−2.15	−1.38	−0.77	−.423	−.255	−0.168	−.136	−.121	−.015
.916	.603	.313	.189	.122	.067	.085	.063	.022
−.538	−.602	.064	−.118	−.137	.019	−.0599	−.074	.014
.995	1.107	−.112	.230	.256	−.026	.126	.146	−.020
.156	.135	.021	.037	.033	.004			

For all considered values of ξ the values of $V - X$ are fairly small after $n = 6$ and can be neglected if two or three-place accuracy is wanted. If the comparison series is not used many more terms would be needed to obtain the same accuracy.

Appendix. *Lemma I. For large n, the coefficients $V_n(\xi)$ are given approximately by*

$$W_n(\xi) = (\xi^2 - 1)^{1/2}Q_n'(\xi)/n(n + 1)Q_n(\xi).$$

From reference [3] we find that for large n

$$ps_n^1(\eta;\gamma^2) \approx P_n^1(\eta),$$
$$S_n^{1(4)}(\xi;\gamma) \approx -iQ_n^1(\xi)/K_n^1(\gamma)n(n + 1).$$

Then

$$V_n(\xi) = S_n^{1(4)}(\xi;\gamma)/[(\xi^2 - 1)^{1/2}S_n^{1(4)}(\xi;\gamma)]' \approx Q_n^1(\xi)/[(\xi^2 - 1)^{1/2}Q_n^1(\xi)]'. \tag{8}$$

Now $Q_n^1(\xi) = (\xi^2 - 1)^{1/2}\, dQ_n/d\xi$ and from the differential equation satisfied by the Q_n we find $[(\xi^2 - 1)^{1/2}\, Q_n^1(\xi)]' = n(n+1)Q_n$. If we make these substitutions in (8) we get the desired result.

Lemma II. For large n,

$$Q_n(\xi) \approx (u/\sinh u)^{1/2} K_0((n + 1/2)u).$$

with $\xi \cosh u$ and $K_0(w)$ the modified Bessel function of the third kind or the Hankel function of imaginary argument, $H_0(iw)$. This result follows immediately from the results of reference [8].

Lemma III. For large n

$$(\xi^2 - 1)^{1/2} Q_n'(\xi)/Q_n(\xi) \approx -(n + 1/2)K_1(w)/K_0(w) + 1/2u - \tfrac{1}{2} \coth u$$

where $w = (n + 1/2)u$ and $K_1(w)$ is again a Hankel function of imaginary argument.

This result follows directly from Lemma II.

Lemma IV. For small w, $K_1(w)/K_0(w) = -w^{-1}[\gamma + \ln w/2]^{-1}$, approximately, where $\gamma = 0.5772$ is Euler's constant; for large w, $K_1(w)/K_0(w) = 1 + 1/2w - 1/8w^2$, approximately.

The first result follows from the definition of $K_0(w)$ and $K_1(w)$ in terms of the modified Bessel functions $I_0(w)$ and $I_1(w)$ together with the approximations, $I_0(w) \approx 1$, $I_1(w) \approx w/2$. The second result follows from the asymptotic expansion of $K_n(w)$ which can be found, e.g. in reference [9].

References

1. H. Myers, *Radiation patterns of the prolate spheroidal antenna*, Trans. I. R. E. **AP-4**, 58–64 (1956)
2. C. P. Wells, *The prolate spheroidal antenna: Current and impedance*, Trans. I. R. E. **AP-6**, 125–128 (1958)
3. J. Meixner and F. W. Schafke, *Mathieusche Funktionen und Spharoidfunktionen*, Springer Verlag, Berlin, 1954
4. C. Flammer, *Spheroidal wave functions*, Standord University Press, Stanford, Calif., 1957
5. J. Meixner and W. Kloepfer, *Theorie der ebenen Ringspalt-antenne*, Z. Physik **3**, 171–178 (1951)
6. G. N. Watson, *Notes on generating functions of polynomials, III, Polynomials of Legendre and Gegenbauer*, J. Lond. Math. Soc. **8**, 289–292 (1933)
7. K. Knopp, *Theory and application of infinite series*, Blackie and Sons, London, 1928
8. G. Szegö, *Entwicklungen der Legendreschen Funktionen*, Proc. Lond. Math. Soc. **36**, 427–450 (1933)
9. A. Erdélyi et al., *Higher transcendental functions* vol. 2, McGraw-Hill, New York, 1953

BOOK REVIEWS

(*Continued from p. 262*)

and stochastic type) are discussed in some detail. The paper closes with the solution of a problem arising from consideration of a miniature mathematical model of the interdependent steel and automobile industries. This leads to a bottleneck problem involving the allocating of steel resources over a given period of time to the production of (a) additional steel itself, (b) additional steel-producing facilities, and (c) automobiles, where the objective is to maximize the total number of automobiles produced during the process.

Even a casual perusal of this volume will impress the reader with the role of the omnipresent variational principle in applied mathematics.

ROBERT KALABA

Boundary layer effects in aerodynamics. International Symposium held at National Physical Laboratory, England, in 1955. Philosophical Library, New York, 1957. $12.00.

This volume contains a complete record of an International Symposium held at the National Physical Laboratory, England, in the Spring of 1955. Nine papers were presented, covering most aspects of the subject of current interest. The Symposium was opened by Professor L. Howarth, with a broad and enlightening survey of boundary layer theory as it stood in 1955.

The first two papers are essentially of a theoretical character. A paper on three dimensional boundary layers by Timman analyses the flow on a general surface in terms of intrinsic coordinates with application to thin wings and yawed elliptic cylinders. Lighthill and Glauert investigate laminar boundary layer flow on a long thin cylinder and obtain solutions valid near the nose and a long way downstream. In a paper on the stability of boundary layer flow on a rotating disk experimental work by Gregory and Walker is compared with theoretical analysis by Stuart. In many respects the two approaches show good agreement and discrepancies which do arise are satisfactorily explained.

An analysis of boundary layer transition by Schubauer and Klebanoff lends strong support to the turbulence spot theory of Emmons. Küchemann considers viscosity effects on swept-back wings with emphasis on separation and the growth of tip vortex sheets. Pankhurst describes various methods of boundary layer control including the use of suction on thick wings.

Young and Kirkby calculate the profile drag on supersonic wings of biconvex and double wedge sections. The symposium ends with two long papers on Shock Wave Boundary Layer Interaction. Holder and Gadd review the work done on this problem on the flat plate and estimate the influence of the effect on the calculation of base drag. The second paper by Pearcey is concerned with Turbulent Boundary Layers on Transonic Airfoils.

Most of these papers have already been published separately but the collected proceedings also contain accounts of the discussion following each paper. This is invariably interesting and in many cases, quite extended.

M. HOLT

Integral equations and their applications to certain problems in mechanics, mathematical physics and technology. By S. G. Mikhlin. Pergamon Press, New York, London, Paris, Los Angeles, 1957. xii + 338 pp. $12.50.

This is a useful addition to the literature in English on integral equations and their applications to continuum mechanics. The first part of the book, which comprises approximately two-fifths of the whole, provides a readable account of the fundamental theory of integral equations. The classical theory due to Fredholm and the Hilbert-Schmidt theory for equations with symmetric kernels are treated in the first two chapters, and the third chapter is devoted to singular integral equations. Methods for approximate solution of the equations are given together with numerical examples. For the most part detailed proofs are given for the theorems used to develop the theory.

(*Continued on p. 284*)

APPROXIMATE DISTRIBUTIONS OF NOISE POWER MEASUREMENTS*

BY

WALTER FREIBERGER AND ULF GRENANDER

Brown University

Summary. The frequency functions of certain spectral estimates are studied analytically and numerically. An approximation is obtained for the case of a Poisson weight function and compared to the true distribution. The eigenvalues of products of Toeplitz matrices play a crucial role in the sampling theory of quadratic forms; an approximation to their distribution is discussed and its accuracy studied numerically. This leads to approximate probability densities which are thought to be valid for moderate or even small sample sizes.

1. Introduction. Let x_t be stationary Gaussian noise with the power spectral density $f(\lambda)$. The parameter t takes on integral values only, $t = \cdots -1, 0, 1, \cdots$. Given a finite sample $x_1, x_2, \cdots, x_n$ we want to estimate $f(\lambda)$, using a quadratic form Q of the type

$$Q = \frac{1}{n} \sum_{\nu,\mu=1}^{n} w_{\nu-\mu} x_\nu x_\mu , \tag{1}$$

where

$$w_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\lambda} w(\lambda)\, d\lambda. \tag{2}$$

Here the weight function $w(\lambda)$ defines the spectral window and satisfies

$$\begin{aligned} &1) \qquad w(\lambda) \geq 0 \\ &2) \quad \frac{1}{2\pi} \int_{-\pi}^{\pi} w(\lambda)\, d\lambda = 1. \end{aligned} \tag{3}$$

If the sample size n is large, it is known that Q has an asymptotically Gaussian distribution with parameters that can be expressed simply in terms of $f(\lambda)$ and $w(\lambda)$; see Grenander and Rosenblatt (1957), p. 134 [1]. On this basis a choice between various *a priori* possible estimates can be rationally made, in accordance with the well-known relation between the band-width and variance of a spectral estimate.

A considerable number of particular weight functions has been suggested in the statistical literature. The usual procedure is to choose an even, periodic weight function $v(\lambda)$, subject to the above conditions (3), and to form

$$w(\lambda) = \tfrac{1}{2}[v(\lambda - \lambda_0) + v(\lambda + \lambda_0)], \tag{4}$$

where λ_0 is the frequency at which we want to determine $f(\lambda_0)$. If the Fourier coefficients of $v(\lambda)$ are denoted by v_k then clearly $w_k = v_k \cos k\lambda_0$.

---

*Received October 22, 1958. The work described here has been partly supported by the U. S. Army Signal Corps, under Contract DA-SC-78130.

We will be concerned with two particular cases. The first is the rectangular weight function

$$v(\lambda) = \begin{cases} \frac{\pi}{h}, & |\lambda| < h \\ 0, & |\lambda| > h. \end{cases} \tag{5}$$

It is of course desirable to keep the bandwidth parameter h small, but it must not be chosen so small that the variance becomes excessively large.

The other is the Poisson weight function

$$v(\lambda) = \frac{1-\rho^2}{|1-\rho e^{i\lambda}|^2} = \frac{1-\rho^2}{1-2\rho\cos\lambda+\rho^2}, \qquad 0 \le \rho < 1, \tag{6}$$

where values of ρ close to one correspond to narrow spectral windows.

The Fourier coefficients v_k are in the first case $\sin kh/kh$ and in the second $\rho^{|k|}$. Both these windows are convenient to deal with and there are reasons to believe they have good sampling properties.

The distribution theory for spectral estimates is fairly complicated. It is true that one can write down immediately the characteristic function $\varphi(z)$ of Q as

$$\varphi(z) = \prod_{\nu=1}^{n} \left(1 - 2\frac{iz\lambda_\nu}{n}\right)^{-1/2}, \tag{7}$$

where the λ_ν are the eigenvalues (always real) of the matrix product RW. Here R is the covariance matrix

$$\begin{aligned} R &= \{r_{\nu-\mu}; \quad \nu, \mu = 1, 2, \cdots, n\} \\ &= \left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(\nu-\mu)\lambda} f(\lambda)\, d\lambda; \quad \nu, \mu = 1, 2, \cdots, n\right\} \end{aligned} \tag{8}$$

and

$$W = \{w_{\nu-\mu}; \quad \nu, \mu = 1, 2, \cdots n\}. \tag{9}$$

The expression (7) is of little immediate practical use since (a) the eigenvalues are in general very difficult to obtain numerically and (b) it is very hard to get the frequency function $g(x)$ of Q by numerical Fourier inversion. Therefore, one is almost forced to resort to approximations, of which the most widely discussed is the Gaussian one mentioned above.

Recently attention has been focused on the need for sharper approximations, valid for moderate sample sizes. One such approximation can be obtained by applying known results from the theory of Toeplitz forms. Indeed, one knows that the λ_ν behave distributionwise approximately as the values of the function $w(\lambda)\, f(\lambda)$, *viz.*

$$\lim_{n\to\infty} \frac{1}{n}\{\text{number of } \lambda_\nu \le \mu\} = \frac{1}{2\pi}\,\text{meas}\,\{\lambda \mid f(\lambda)w(\lambda) \le \mu\}; \tag{10}$$

see Grenander and Szegö (1958), Chap. 8 [2]. This implies that we have approximately

$$\log\varphi(z) = -\tfrac{1}{2}\sum_{\nu=1}^{n} \log\left(1 - \frac{2iz\lambda_\nu}{n}\right) \cong -\frac{n}{4\pi}\int_{-\pi}^{\pi} \log\left[1 - \frac{2izf(\lambda)w(\lambda)}{n}\right] d\lambda. \tag{11}$$

Before entering into a discussion of how (11) can be used to determine the approximate frequency function of Q, one important observation will be made. It is clear that a spectral estimate will be useful only if it resolves the spectrum well enough so that $w(\lambda)$ is peaked around the frequency of interest λ_0 . If the spectral window is narrow, measured in terms of a typical frequency of the stochastic process, then the values of the distributions of the two functions $f(\lambda)\ w(x)$ and $f(\lambda_0)\ w(\lambda)$ are almost identical, so that little is changed in (11) if we substitute the second function for the first one.

Hence the stochastic variable $Q/f(\lambda_0)$ will have approximately the characteristic function

$$E \exp [izQ/f(\lambda_0)] \cong \varphi_n(z) = \exp \left\{ -\frac{n}{4\pi} \int_{-\pi}^{\pi} \log \left[1 - \frac{2izw(\lambda)}{n} \right] d\lambda \right\}. \tag{12}$$

This distribution apparently does not depend upon the (unknown) spectral density $f(\lambda)$, so that we can use it to construct approximate confidence limits $f_1^*(\lambda_0)$ and $f_2^*(\lambda_0)$. Indeed, $f_i^*(\lambda_0) = Q/x_i$ with

$$\int_{x_1}^{x_2} g_n(x)\, dx = p, \tag{13}$$

where p is the confidence coefficient and $g_n(x)$ is the frequency function corresponding to $\varphi_n(z)$.

A parallel investigation, Grenander, Pollak and Slepian (1959) [3], discussed the approximation (11) for $f(\lambda) \equiv 1$. The approximation mentioned in the last paragraph does not seem to have been studied numerically in the literature. This will be done for a few cases in the following section.

2. Numerical studies. Let us introduce the eigenvalues λ_k' $(k = 1, 2, \cdots, n)$ of the matrix $f(\lambda_0)W$; as mentioned above, these eigenvalues can be expected to have about the same distribution as λ_k . In studying this approximation numerically we are constrained by the limitations of our computing facilities. To avoid excessive computing effort, we have chosen $n = 20$. This number may not correspond to situations met frequently in practice, but it is hoped that the information gained will be of some general interest.

The spectral density was chosen as

$$f(\lambda) = 102 + 22 \cos 2\lambda + 20 \cos 4\lambda, \tag{14}$$

a graph of which is shown in Fig. 1(1). The value of $f(\lambda)$ at $\lambda_0 = 0$ is $f(0) = 144$; the Poisson weight function (6) was used with $\rho = .7$, .9 and .95. The eigenvalues λ_k and λ_k' are given in Table 1.

The statistical distribution of the noise-power estimates will depend mainly upon the largest eigenvalues and the table shows that the error in these is of the order of a few per cent. We have also computed the exact frequency function $g(x)$ and compared it with the frequency function corresponding to the eigenvalues λ_k' ; see Figs. 2–4.

The frequency functions have been computed from the expression

$$g(x) = \frac{1}{\pi} \sum_{\nu=1}^{10} (-1)^{\nu+1} \int_{(2\lambda_{2\nu-1})^{-1}}^{(2\lambda_{2\nu})^{-1}} e^{-ux} \prod_{j=1}^{20} (1 - 2y\lambda_j)^{-1/2}\, dy \tag{15}$$

and the corresponding formula with the λ_ν replaced by the λ_ν' . This expression was given by Slepian (1958) [4]. To eliminate singularities at the ends of the intervals of

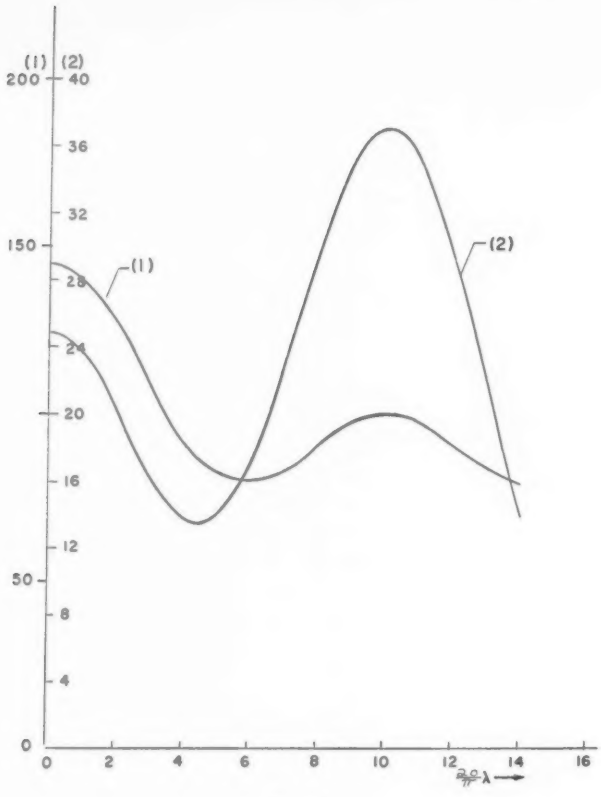

(1)
(2)
200
40
36
32
150
(1)
(2)
28
24
20
16
12
50
8
4
0
0
2
4
6
8
10
12
14
16
$\frac{2\pi}{\nu}\lambda$

FIG. 1

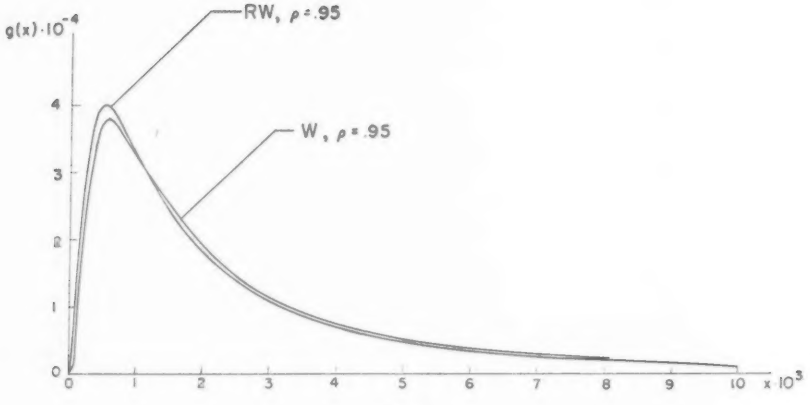

$g(x)\cdot 10^{-4}$
RW, $\rho = .95$
W, $\rho = .95$
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
x 10^3

FIG. 2

TABLE 1

k	$\rho = .95$		$\rho = .9$		$\rho = .7$	
	λ_k	λ_k'	λ_k	λ_k'	λ_k	λ_k'
1	2078.9	2114.6	1596.0	1618.3	723.3	729.6
2	386.9	405.4	552.1	576.0	529.4	547.3
3	124.6	134.1	220.2	237.5	353.4	379.0
4	57.1	64.2	108.8	122.4	232.0	260.9
5	31.6	37.6	62.1	73.9	155.2	185.0
6	19.6	24.8	39.0	49.5	107.0	136.4
7	13.0	17.7	26.2	35.7	76.3	104.3
8	9.2	13.4	18.6	27.2	56.2	82.7
9	6.8	10.7	13.8	21.6	42.8	67.5
10	5.2	8.8	10.7	17.7	33.9	56.6
11	4.3	7.3	8.7	15.0	27.9	48.5
12	3.6	6.3	7.3	13.0	23.9	42.5
13	3.2	5.6	6.5	11.5	21.2	37.9
14	2.9	5.0	5.9	10.4	19.5	34.3
15	2.7	4.6	5.6	9.5	18.5	31.5
16	2.6	4.3	5.4	8.9	18.0	29.5
17	2.6	4.0	5.3	8.3	17.7	27.9
18	2.6	3.9	5.3	8.1	17.6	26.8
19	2.6	3.7	5.3	7.8	17.6	26.1
20	2.6	3.7	5.3	7.6	17.6	25.3

integration, a change of variable

$$y = c_\nu \cos \pi x + d_\nu ,$$

with

$$c_\nu = \tfrac{1}{2}(y_{2\nu+1} - y_{2\nu}), \qquad d_\nu = \tfrac{1}{2}(y_{2\nu+1} + y_{2\nu}), \qquad y_\nu = \frac{1}{2\lambda_\nu}$$

is effected and the resulting integrals are evaluated by a Newton-Coates five-point formula.

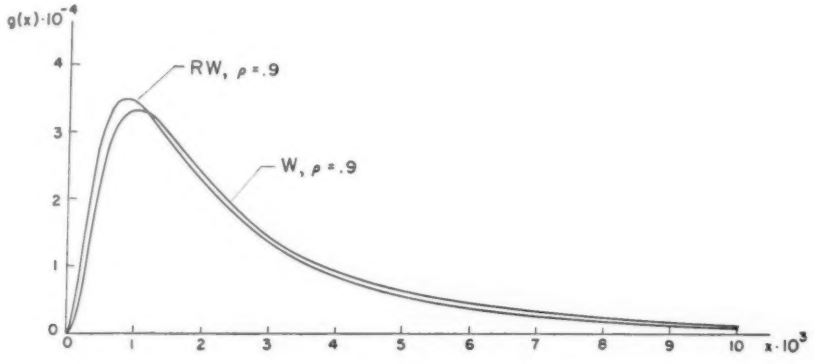


FIG. 3

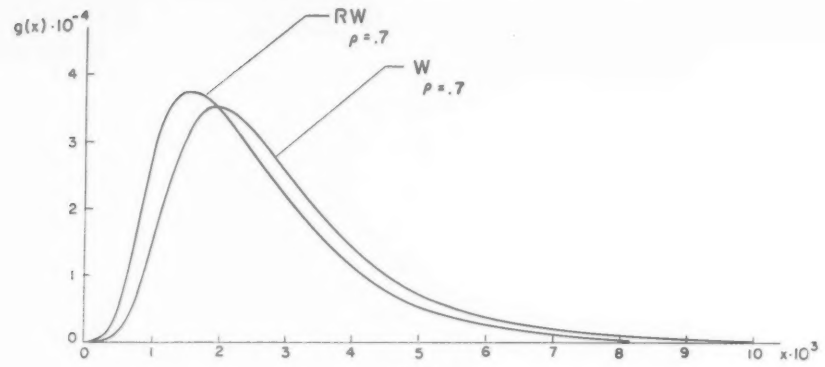


FIG. 4

Because of the proximity of the eigenvalues, which is characteristic for matrices of this type, the classical iteration procedure broke down. Instead, the Jacobi diagonalisation method was modified to apply to (non-symmetric) matrices which are products of two symmetric matrices; see Kalker (1958) [5]. The computations were carried out on an IBM 650 computer and took of the order of a few hours for the determination of the eigenvalues and of each frequency function.

It is obvious from the graphs given below that the approximation works well for the narrow windows $\rho = .95, .90, .70$. If the weight function is flat one will not expect the approximation to be adequate and in the cases $\rho = .5$ and .3 we give only the λ_k' in Table 2 and the corresponding frequency functions in Figs. 5 and 6. This table may be of interest to a reader who wishes to investigate wide spectral windows.

TABLE 2

k	$\rho = .5$	$\rho = .3$
	λ_k'	λ_k'
1	416.4	264.1
2	375.5	254.6
3	322.6	240.3
4	269.1	223.1
5	221.9	204.6
6	183.3	186.2
7	152.6	168.9
8	128.7	153.4
9	110.2	139.4
10	95.6	127.4
11	84.2	117.2
12	75.3	108.4
13	68.3	101.1
14	62.8	95.0
15	58.3	90.0
16	54.9	86.0
17	52.3	82.8
18	50.4	80.5
19	49.0	78.8
20	48.2	77.9

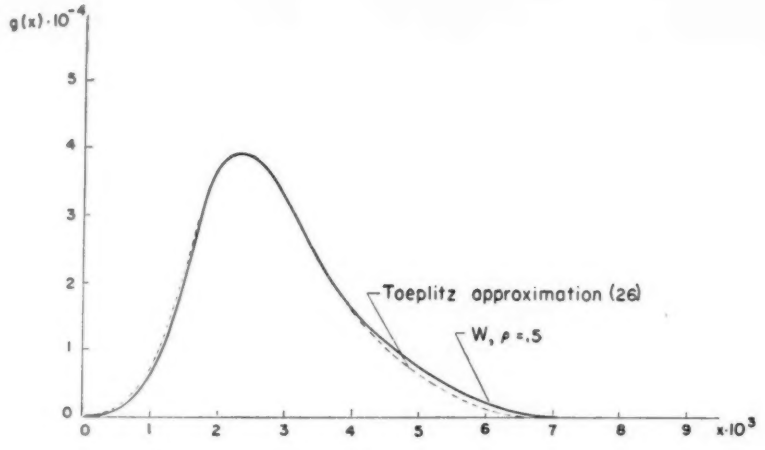


FIG. 5

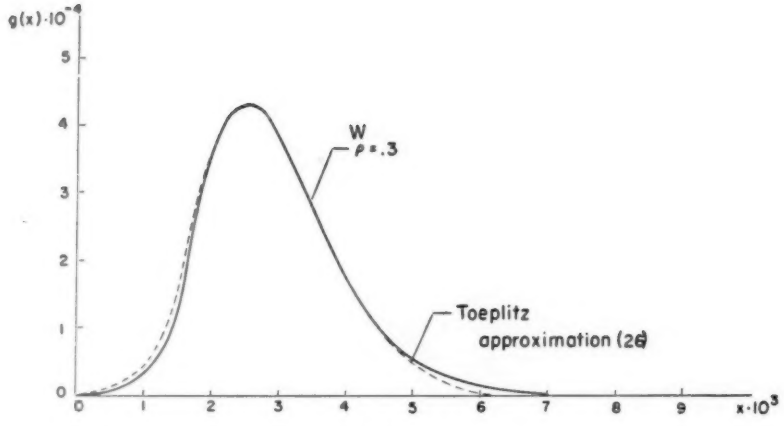


FIG. 6

For the rectangular window (5) the analogous quantities have been computed for the spectral density

$$f(\lambda) = 19 + 6\cos\lambda - 10\cos 2\lambda + 2\cos 3\lambda + 8\cos 4\lambda \tag{16}$$

the graph of which is shown in Fig. 1(2). The value of $f(\lambda)$ at the center $\lambda_0 = 3/10\,\pi$ of the windows is $f(\lambda_0) = 17.2$. The weight function (5) was used with

$$h = \frac{\pi}{40}, \quad \frac{\pi}{20}, \quad 3\frac{\pi}{40}, \quad \frac{\pi}{8}, \quad \frac{\pi}{4}.$$

For the first three of these values we have Table 3.

TABLE 3

k	$h = \pi/40$		$h = \pi/20$		$h = 3\pi/40$	
	λ_k	λ_k'	λ_k	λ_k'	λ_k	λ_k'
1	176.2	161.2	147.5	135.3	118.9	107.2
2	166.6	160.9	140.9	134.7	113.7	106.7
3	13.8	10.9	44.0	35.2	68.2	56.5
4	11.6	10.7	37.2	34.9	58.5	56.2
5	.175	.1	2.7	1.9	12.1	8.7
6	.126	.1	2.0	1.7	9.0	8.0
7	0	0	.1	0	.6	0
8	0	0	0	0	.3	0
9	↓	↓	0	0	0	0
10			↓	↓	0	0
11						
12					↓	↓

While the approximation still seems adequate it is interesting to note that the relative errors of the largest eigenvalues are now somewhat larger than for the Poisson weight function. The explanation for this may be found in the local behavior of the spectral densities at the frequencies of interest.

The resulting frequency functions are represented in Figs. 7–9. We also give the values of λ_k' for $h = \pi/8$ and $\pi/4$, in Table 4.

TABLE 4

k	$h = \pi/8$	$h = \pi/4$
	λ_k'	λ_k'
1	46.0	30.9
2	45.7	26.6
3	29.7	23.7
4	27.8	21.1
5	17.0	19.6
6	16.0	18.9
7	14.9	18.4
8	14.9	18.4
9	14.9	18.2
10	14.9	18.2
11	14.9	16.6
12	14.9	16.6
13	14.8	16.5
14	14.6	16.4
15	14.4	16.2
16	14.3	15.8
17	11.6	14.3
18	10.6	10.4
19	4.8	3.9
20	3.3	3.5

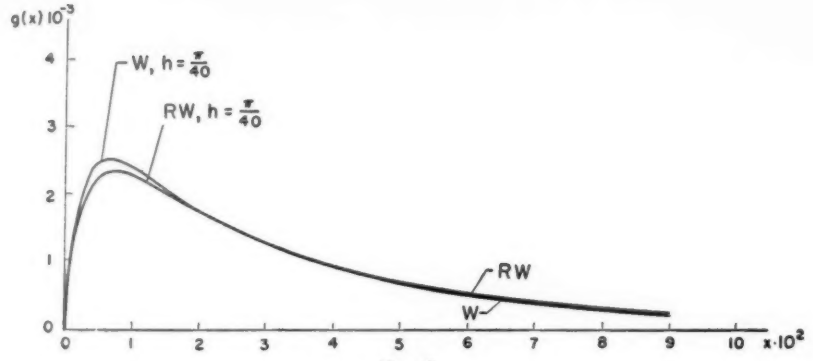


FIG. 7

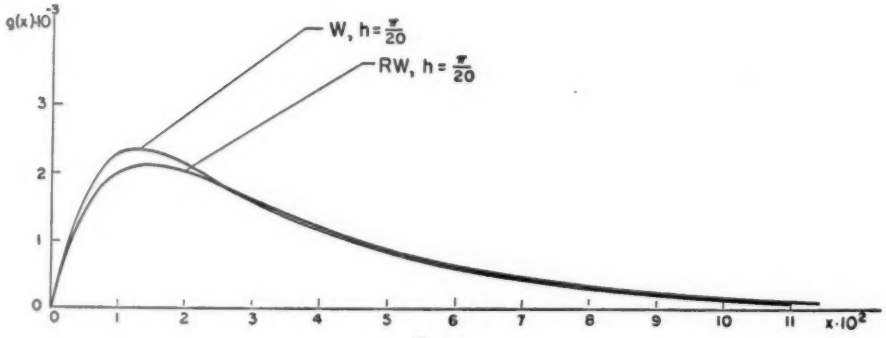


FIG. 8

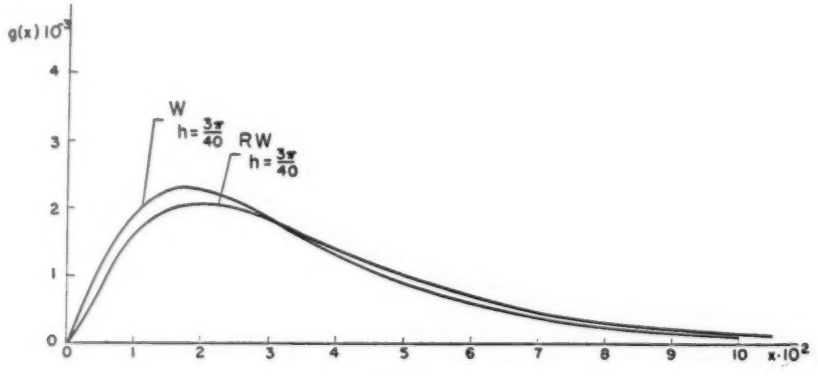


FIG. 9

Similar observations to those made in the case of the Poisson weight function seem to apply here.

3. Some closed-form approximations. We have seen in the last section that if the spectral window is sufficiently narrow, little accuracy is lost if we regard $f(\lambda)$ constant

throughout the interval $(-\pi, \pi)$. If this is done, and if $f(\lambda)$ is normed to 1, we can use Eq. (12) to give us the approximate expression for the characteristic function

$$\varphi_a(z) = \exp\left[-\frac{n}{2}\psi(z)\right],$$

where

$$\psi(z) = \frac{1}{2\pi}\int_0^{2\pi} \log\left[1 - \frac{2izw(\lambda)}{n}\right] d\lambda. \tag{17}$$

For the case of a rectangular window this equation gives us the characteristic function of the well-known type III distribution with

$$\psi(z) = \frac{h}{\pi}\log\left[1 - \frac{2\pi iz}{nh}\right]$$

so that

$$\varphi_a(z) = \left[1 - \frac{2\pi iz}{nh}\right]^{-nh/2\pi}.$$

This corresponds to the frequency function

$$g_a(x) = \left(\frac{nh}{\pi}\right)^{hn/2\pi} \frac{1}{2^{hn/2\pi}\Gamma(hn/2\pi)}\, x^{hn/2\pi - 1} \exp\left[-(xnh/2\pi)\right], \tag{18}$$

an approximation which may in essence be contained in the work of S. O. Rice.

The Toeplitz approximation for a Poisson spectral window can also be expressed in closed form. Indeed, the function

$$\exp\psi(z) = \exp\frac{1}{2\pi}\int_0^{2\pi} \log\left[1 - \frac{2iz}{n}\frac{1-\rho^2}{|\,1-\rho e^{i\lambda}\,|^2}\right] d\lambda \tag{19}$$

is analytic in z in the plane cut from

$$z = -\frac{n}{2}\frac{1-\rho}{1+\rho}i \quad \text{to} \quad z = -\frac{n}{2}\frac{1+\rho}{1-\rho}i, \qquad \text{see Fig. 10.}$$

Noticing that

$$|\operatorname{Re}\psi(z)| \leq \tfrac{1}{2}\log\left[1 + \frac{4z^2}{n^2}\left(\frac{1+\rho}{1-\rho}\right)^2\right], \tag{20}$$

so that $|\varphi_a(z)|$ is dominated by a power of z, we can change that contour of integration in the Fourier inversion formula

$$g_a(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixz}\varphi(z)\,dz \tag{21}$$

to give us

$$g_a(x) = \frac{1}{2\pi}\oint e^{-ixz}\varphi(z)\,dz. \tag{22}$$

However

$$\psi(z) = \frac{1}{2\pi}\int_0^{2\pi}\left\{\log\left[|\,1-\rho e^{i\lambda}\,|^2 - \frac{2iz}{n}(1-\rho^2)\right] - \log|\,1-\rho e^{i\lambda}\,|^2\right\} d\lambda.$$

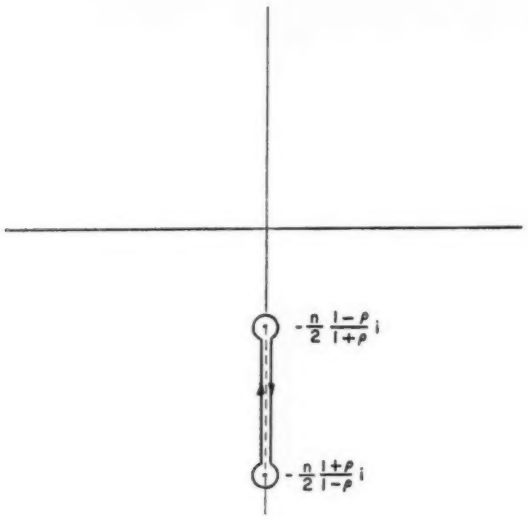


FIG. 10. z-plane.

The last term does not contribute anything to the integral and putting $z = -iw$ we get

$$\psi(z) = \frac{1}{2\pi}\int_0^{2\pi} \log\,[B \mid 1 - Re^{i\lambda} \mid^2]\, d\lambda,$$

where

$$B \mid 1 - Re^{i\lambda} \mid^2 = \mid 1 - \rho e^{i\lambda} \mid^2 - \frac{2v}{n}(1 - \rho^2)$$

which determines the constants

$$B = \frac{\rho}{R}$$

and

$$R = \frac{1 + \rho^2 - \frac{2v}{n}(1 - \rho^2) \pm i\left\{4\rho^2 - \left[1 + \rho^2 - \frac{2v}{n}(1 - \rho^2)\right]^2\right\}^{1/2}}{2\rho}. \tag{23}$$

Introducing the angle θ by

$$R = \cos\theta - i\sin\theta = e^{-i\theta}, \qquad \cos\theta = \mathrm{Re}\,R = \frac{1 + \rho^2 - \frac{2v}{n}(1 - \rho^2)}{2\rho},$$

we get $\psi(z) = \log B$ so that

$$\varphi_a(z) = B^{-n/2} = \left(\frac{R}{\rho}\right)^{n/2} = \frac{e^{-in\theta/2}}{\rho^{n/2}}. \tag{24}$$

Inserting (24) into (22) we get

$$g_\alpha(x) = -\frac{i}{2\pi\rho^m}\oint e^{-vx-im\theta}\,dv,$$

where we have supposed the sample size n to be even, $n = 2m$. The above expression reduces to

$$g_\alpha(x) = \frac{2m}{\pi\rho^{m-1}(1-\rho^2)}\int_0^\pi e^{-vx}\sin m\theta\sin\theta\,d\theta$$

$$= \frac{2me^{-\alpha x}}{\pi\rho^{m-1}(1-\rho^2)}\int_0^\pi e^{-\beta x\cos\theta}\sin m\theta\sin\theta\,d\theta$$

with

$$\left.\begin{aligned}\alpha &= m\frac{1+\rho^2}{1-\rho^2}\\ \beta &= -\frac{2m\rho}{1-\rho^2}\end{aligned}\right\}. \tag{25}$$

Hence

$$\begin{aligned} g_\alpha(x) &= \frac{me^{-\alpha x}}{\pi\rho^{m-1}(1-\rho^2)}\int_0^\pi e^{-\beta x\cos\theta}\left[\cos(m-1)\theta-\cos(m+1)\theta\right]d\theta\\ &= -\frac{me^{-\alpha x}i^{m+1}}{\rho^{m-1}(1-\rho^2)}\left[J_{m-1}(i\beta x)+J_{m+1}(i\beta x)\right]\\ &= -\frac{me^{-\alpha x}i^{m+1}}{\rho^{m-1}(1-\rho^2)}\left[i^{m-1}I_{m-1}(\beta x)+i^{m+1}I_{m+1}(\beta x)\right]\\ &= (-1)^m\frac{me^{-\alpha x}}{\rho^{m-1}(1-\rho^2)}\left[I_{m+1}(\beta x)-I_{m-1}(\beta x)\right]\\ &= \frac{2m^2e^{-\alpha x}(-1)^{m+1}}{\rho^{m-1}(1-\rho^2)\beta}\frac{1}{x}I_m(\beta x)\\ &= \frac{me^{-\alpha x}(-1)^m}{x\rho^m}I_m(\beta x). \end{aligned} \tag{26}$$

As a numerical test of the accuracy of this Toeplitz approximation we have plotted (26) and the corresponding frequency functions of the exact distributions for $\rho = .5$ and $\rho = .3$ in Figs. 5 and 6. Because of the relatively small sample size used in the computations, the approximation (26) was thought to be adequate only for small values ρ (wide spectral windows); it turned out, however, to be surprisingly close, as seen in the figures, so that its validity may well extend to somewhat narrower windows than expected.

4. Discussion of the approximations. We have examined the behavior of the eigenvalues of the matrix product RW and compared it to that of $f(\lambda_0)W$. Because of the limited scope of this numerical study, our conclusions are necessarily of a tentative nature. It seems that the approximation is better the larger the sample size, the narrower the spectral window and the flatter the spectral density. Obviously, the spectral window has to be taken narrow enough not to smooth out too much of the large scale structure

of the spectrum. If the time constant corresponding to this spectral window is only a fraction of n we would expect this approximation to be satisfactory.

The Toeplitz approximation (26) seems to be valid if the time constant $\rho/1 - \rho$ is small compared to n.

Acknowledgment. We would like to thank Miss Cornelia M. Kalkman for her assistance with the numerical work.

References

1. U. Grenander and M. Rosenblatt, *Statistical analysis of stationary time series*, John Wiley & Sons, New York, 1957
2. U. Grenander and G. Szegö, *Toeplitz forms and their applications*, University of California Press, Berkeley and Los Angeles, 1958
3. U. Grenander, H. O. Pollak and D. Slepian, *The distribution of quadratic forms in normal variates: A small sample theory with applications to spectral analysis*, (to be published)
4. D. Slepian, *Fluctuations of random noise power*, The Bell System Tech. J. **37**, 163–184 (Jan. 1958)
5. J. J. Kalker, *IBM 650 programs for matrix computations based on Jacobi's method*, Brown University Rept. DA-SC-78130/1, 1958

BOOK REVIEWS

(*Continued from p. 270*)

The second part of the book contains a variety of applications, almost all of which are two-dimensional problems. As is to be expected and desired in a work of this nature, the emphasis is on the method of application of integral equations to obtain the solution, so that the working is carried only to the point where detailed results can be found by straightforward calculations. Discussions are included of the integral equation approach to the solution of the wave equation and the heat conduction equation but most of the examples are taken from Russian work in the plane theory of elasticity and hydrodynamics.

The translator deserves thanks for a translation which very seldom jars. However, the reader may be somewhat startled to find that Hooke has lost his final "e" due to the strain of being re-transliterated into English, and one may seek in vain for references numbered higher than thirty-five.

R. T. Shield

Vibration and impact. By Ralph Burton. Addison-Wesley Publishing Co., Reading, Mass. x + 310 pp. $8.50.

The author states that this book is intended for senior undergraduate, and for beginning graduate students, and the objectives of the text are listed as treatments of (1) "the natural frequency" (free vibrations without damping), (2) "forces which tend to suppress vibrations" (free vibrations with damping), (3) "periodic forced and transient forced vibrations", and (4) "self-excited vibrations". There are chapters on free vibration, vibratory systems commonly found in machinery, damping, impact, nonlinear vibrations, measuring instruments and analogs, numerical computation of natural frequencies (of multi-modal systems), waves, vibrating beams and related subjects, (the related subjects including such items as plate and membrane vibrations, and damping in a system having two lumped masses, two springs and one viscous damper), and chapters on the analysis of control systems, and fatigue. Nearly every chapter contains some examples and is followed by a number of problems.

In view of the title of the text, the reader may expect to find a book devoted in equal parts to vibrations and to impact; he will be disappointed. There is one chapter of approximately twenty pages dealing with elementary aspects of the response of rigid bodies, elastically supported, to non-periodic, time-dependent forces. In another chapter, there is a short qualitative discussion of wave propagation, wave interaction, and wave reflection from discontinuities, consisting of largely verbal results deduced from heuristic arguments; however, this latter material does not add greatly to the study of impact on elastic bodies.

The instructor who intends to use this book as a text in senior undergraduate, or in graduate, courses must expect students who have no background whatever in ordinary differential equations but a fair preparation in partial differential equations, for there are quite lengthy derivations of the solutions of linear, homogeneous equations with constant coefficients, while the solution to the one-dimensional wave equation (the only partial differential equation in the book) is given without any derivation whatever.

To the reviewer, the book appears disconnected, badly arranged and replete with errors of commission and omission. To list some of the former, the author says of the equation $-kx = m\ddot{x}$ that "unfortunately, we cannot solve this equation by direct integration". The "perturbation method" is illustrated by the example of the pendulum, and the author believes that this method consists of annihilating the nonlinear terms by amputation; thus the "perturbation solution" of $\ddot{\theta} + (g/l) \sin \theta = 0$ is obtained by solving $\ddot{\theta} + (g/l)\theta = 0$. The author notes that "Fourier analysis cannot be applied to the solution of nonlinear systems under periodic nonharmonic forcing. . . . Any solution based upon superposition of solutions is invalid . . .". The criterion for stability of vibrating systems is said to be the same as that for static stability—i.e., that the system ". . . should return to its initial configuration after a small disturbance". Of Raleigh's method, it is said that "the more boundary conditions the assumed solution meets, the better will be the estimated frequency". It is also said that "It can be shown that

(*Continued on p. 298*)

FORCED HEAT CONVECTION IN LAMINAR FLOW THROUGH RECTANGULAR DUCTS*

BY

S. C. R. DENNIS, A. McD. MERCER AND G. POOTS

The Queen's University of Belfast

Introduction. In this paper we consider the problem of finding the heat transfer to the wall of a duct which has a rectangular cross-section and through which a hot viscous fluid passes in steady established laminar motion. We shall make the usual simplifying assumptions that the thermal properties of the fluid are independent of temperature, that liquids are incompressible and that gases obey the perfect gas law. The first is strictly true only for small heat input and, of course, the assumption of established motion ignores the hydrodynamical boundary layer in the inlet. The problem is of engineering interest since, in many applications of gas-flow heat exchangers, flow passages are used which have small cross-section and a high ratio of surface area to core volume, so that the Reynolds number is often small enough for laminar flow to exist. Practical cross-sections can often be approximated by simple geometrical shapes and theoretical correlations of heat transfer with cross-section are of value in reducing the amount of practical test data required.

The rectangular cross-section gives rise to an essentially three-dimensional temperature distribution and has therefore received less attention than those involving two-dimensional distributions, such as the circle and the case of infinite parallel walls. Some results have been given by Clark and Kays (1953) [1] by considering conditions far enough from the thermal inlet to assume a fully developed temperature profile, but this gives only asymptotic results and no information is obtained on the variation of heat transfer in the thermal entry region. On the other hand experimental data are given regarding this variation and it is therefore of interest to obtain theoretical results taking into account the undeveloped temperature profile. This is the object of this paper although it is hoped that the numerical analysis of the governing partial differential equation, which occurs in wider fields, will also be of interest.

Governing equations and basic thermal quantities. We consider a duct whose axis is the ζ-axis of rectangular coordinates (ξ, η, ζ) and whose cross-section perimeter is, in general, the curve $C(\xi, \eta) = 0$. The constant cross-sectional area is A and the length of the perimeter is S. In customary notation the velocity field is (u, v, w), but for steady established laminar motion under a constant pressure gradient P/L we have $u = v = 0$ and $w \equiv w(\xi, \eta)$ where

$$\frac{\partial^2 w}{\partial \xi^2} + \frac{\partial^2 w}{\partial \eta^2} = -\frac{P}{\mu L}, \tag{1}$$

with $w = 0$ on C. The energy equation governing the temperature $T(\xi, \eta, \zeta)$ of the fluid is, subject to the stated assumptions,

$$\kappa\left(\frac{\partial^2 T}{\partial \xi^2} + \frac{\partial^2 T}{\partial \eta^2} + \frac{\partial^2 T}{\partial \zeta^2}\right) = w \frac{\partial T}{\partial \zeta}, \tag{2}$$

*Received Nov. 5, 1957; revised manuscript received Sept. 26, 1958.

where κ is the thermometric conductivity, and second order terms, such as that due to internal heat generation, are neglected. The origin is at the thermal inlet and we suppose that the fluid enters $\zeta \geq 0$ with constant temperature T_0 . We introduce dimensionless coordinates $x = \xi/d$, $y = \eta/d$ and $z = \zeta/d$ Pé where $d = 4A/S$ and Pé is the Péclet number, equal to the product of the Reynolds and Prandtl numbers, that is, dw_0/κ. The Reynolds number $dw_0\rho/\mu$ is based on the mean velocity w_0 , that is, the ratio of total flow to cross-sectional area. It now appears that the ratio of axial conduction $\kappa\partial^2 T/\partial\zeta^2$ to the axial convection term $w\partial T/\partial\zeta$ is of order $(1/\text{Pé})^2$, so that axial conduction may be neglected for large enough Pé. This can lead to discrepancies in the special case of low Reynolds number flow of low Prandtl number fluids, such as certain liquid metals, but for water, air and high Prandtl number oils it is justified. Equations (1) and (2) then become

$$\nabla_1^2 w = -\frac{d^2}{\mu}\frac{P}{L} \tag{3}$$

and

$$\nabla_1^2 \theta = \frac{w}{w_0}\frac{\partial}{\partial z}\frac{\theta}{}, \tag{4}$$

where

$$\nabla_1^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \quad \text{and} \quad \theta = \frac{T - T_1}{T_0 - T_1},$$

T_1 $(< T_0)$ being a representative temperature associated with the duct wall in the region $\zeta > 0$. The boundary condition for Eq. (3) is that $w = 0$ on the transformed boundary $C'(x, y) = 0$ while that for Eq. (4) depends upon the assumptions that are made. If T_1 is taken as the wall temperature, assumed constant, we have $\theta = 1$ within C' when $z = 0$ and $\theta = 0$ on C' when $z > 0$. There is also another boundary condition in which we interpret T_1 as the temperature of the medium just outside the duct wall, again assumed constant. It has been shown by Hampton [2] that, when heat losses by radiation and natural convection take place from a body at temperature T into surroundings at temperature T_1 , the flux of heat $H(\text{cal/cm}^2/\text{sec})$ is well represented for temperatures from 0 − 400°C. by the empirical formula

$$H = A(T - T_1) + B(T - T_1)^2. \tag{5}$$

The constants A and B depend upon the emissivity E of the body, A varying from 1.96×10^{-4} when $E = 1$ to 1.33×10^{-4} when $E = 0$ and B varying correspondingly from 1.71×10^{-6} to 0.25×10^{-6}. If we identify T with the temperature of the duct wall (assumed ideally to be of negligible thickness so that T is the temperature of the fluid in contact with it) the heat flux to the wall from the fluid is $H = -k\partial T/\partial\nu$, where k is the thermal conductivity of the fluid and ν is the outward drawn normal from the duct wall. Substituting in Eq. (5) and introducing dimensionless quantities we obtain

$$-\frac{k}{d}\frac{\partial}{\partial\nu'}\frac{\theta}{} = A\theta + B(T_0 - T_1)\theta^2, \tag{6}$$

where $\nu' = \nu/d$. Now $\theta < 1$, tending to zero for large z, while B is of order 10^{-2} A so that for small $T_0 - T_1$ (which is implied in the basic assumptions) we may neglect the

second term on the right hand side of Eq. (6). Putting $N = Ad/k$, the complete boundary conditions for θ in this case may therefore be stated as

$$\theta = 1 \text{ within } C' \text{ when } z = 0, \ \partial\theta/\partial\nu' = -N\theta \text{ on } C' \text{ when } z > 0. \tag{7}$$

If N is infinite we get $T = T_1$ at the wall, so that this special case gives the constant wall temperature condition. If $N = 0$ we have the trivial solution $T = T_0$. In practice N must lie somewhere in between these limits.

The solution of Eq. (4) may be written

$$\theta = \sum_{n=1}^{\infty} \alpha_n \Theta_n(x, y) \exp(-\lambda_n z), \qquad (n = 1, 2, 3 \cdots), \tag{8}$$

where Θ_n and λ_n are eigenfunctions and eigenvalues of the membrane equation

$$\nabla_1^2 \Theta + \frac{\lambda w(x, y)}{w_0} \Theta = 0, \tag{9}$$

subject to the boundary condition, deduced from the second of the conditions (7), that $\partial\Theta/\partial\nu' = -N\Theta$ on C'. The theory of this equation is well known and is dealt with, for example, by Courant and Hilbert [3]. Since $w(x, y)$ is positive within C' the eigenvalues λ_n are real and positive and the eigenfunctions form a complete orthogonal set satisfying the property

$$\iint_{D'} w\Theta_m\Theta_n \, dx \, dy = 0, \qquad (m \neq n), \tag{10}$$

where D' is the domain bounded by C'. Each Θ_n has arbitrary amplitude which we choose, for convenience, so that

$$\iint_{D'} w\Theta_n^2 \, dx \, dy = \iint_{D'} w \, dx \, dy. \tag{11}$$

Putting $z = 0$ in Eq. (8) we have from first of the conditions (7) that

$$1 = \sum_{n=1}^{\infty} \alpha_n \Theta_n \tag{12}$$

so that from Eqs. (10) and (11)

$$\alpha_n = \iint_{D'} w\Theta_n \, dx \, dy \Big/ \iint_{D'} w \, dx \, dy. \tag{13}$$

The temperature $\theta(x, y, z)$ is therefore known to any desired accuracy once sufficient Θ_n have been found. Two further thermal quantities are of interest. Experimental measurements are made on the basis of a mean mixed temperature of the fluid, that is, $\theta(x, y, z)$ averaged with respect to the local fluid velocity over any section of the duct. This temperature is a function of z only and its difference between any two sections gives a measure of the heat transferred across the wall between them. Denoting it by T_M then $\Theta_M = (T_M - T_1)/(T_0 - T_1)$ and is given by

$$\begin{aligned} \theta_M(z) &= \iint_{D'} w\theta \, dx \, dy \Big/ \iint_{D'} w \, dx \, dy, \\ &= \sum_{n=1}^{\infty} \alpha_n^2 \exp(-\lambda_n z). \end{aligned} \tag{14}$$

The remaining quantity to be considered is the rate of heat transfer per unit area to the wall of the duct, defined by means of a heat transfer coefficient. If H is the heat flux to a given area of the duct wall, the fundamental equation defining such a coefficient, h, is $H = h\Delta T$, where ΔT is a representative temperature difference. We consider two coefficients, of somewhat different natures, obtained by different choices of ΔT. First taking $\Delta T = T_M - T_1$ we obtain the local coefficient of heat transfer h_L, which measures the local average rate of heat transfer to the duct wall as a function of longitudinal distance down the duct. The total heat transferred to the area between the sections at ζ and $\zeta + d\zeta$ is $dq = h_L S d\zeta(T_M - T_1) = h_L S d\zeta \theta_M (T_0 - T_1)$ and if s is the distance measured along the perimeter of the boundary curve C in an anti-clockwise direction we must also have

$$dq = -k\, d\zeta \int_C \frac{\partial T}{\partial \nu}\, ds = -k(T_0 - T_1)\, d\zeta \int_C \frac{\partial \theta}{\partial \nu}\, ds. \tag{15}$$

We equate these two and introduce the appropriate dimensionless heat transfer coefficient or Nusselt number, defined as $h_L d/k$, so that we obtain for the local Nusselt number

$$Nu(z) = -d \int_{C'} \frac{\partial \theta}{\partial \nu'}\, ds'/S\theta_M\,, \tag{16}$$

where $s' = s/d$. We now eliminate θ in terms of the Θ_n by Eq. (8) and apply Green's theorem to Eq. (9), so that

$$\int_{C'} \frac{\partial \Theta_n}{\partial \nu'}\, ds' = -\frac{\lambda_n}{w_0} \iint_{D'} w\Theta_n\, dx\, dy. \tag{17}$$

Using Eq. (13) and since $\iint_{D'} w\, dx\, dy = A' w_0$, where A' is the dimensionless area within C', we finally obtain

$$Nu(z) = \frac{1}{4\theta_M} \sum_{n=1}^{\infty} \lambda_n \mathcal{B}_n \exp(-\lambda_n z), \tag{18}$$

where $\mathcal{B}_n = \mathcal{A}_n^2$. At large distances down the duct $Nu(z)$ approaches a limiting minimum value. If λ_1 is the smallest of the λ's we have, as $z \to \infty$, that $4\theta_M Nu(z) \sim \lambda_1 \mathcal{B}_1 \exp(-\lambda_1 z)$ and $\theta_M(z) \sim \mathcal{B}_1 \exp(-\lambda_1 z)$ so that $Nu(\infty) = \lambda_1/4$. For experimental measurements a mean coefficient is generally more useful than the local coefficient. This is based on the total heat, q, transferred to the wall between the thermal inlet and the section ζ. Definition of this coefficient again depends upon the choice of ΔT and the one most commonly used is the logarithmic mean temperature difference

$$\Delta T_{ln} = \frac{\Delta T \max - \Delta T \min}{\ln(\Delta T \max) - \ln(\Delta T \min)} = \frac{(T_0 - T_1) - (T_M - T_1)}{\ln(T_0 - T_1) - \ln(T_M - T_1)}.$$

Adopting this definition in the fundamental equation we have $q = h_{ln} S\zeta \Delta T_{ln}$. Now q can either be obtained by integrating Eq. (15) from zero to ζ or, alternatively, it is the heat given up by the fluid in cooling from T_0 to T_M, so that $q = A w_0 \rho C_p (T_0 - T_M)$. Equating these two and introducing dimensionless quantities we have the mean logarithmic Nusselt number, $h_{ln} d/k$, given by

$$Nu'(z) = \frac{1}{4z} \ln\left(\frac{1}{\theta_M}\right). \tag{19}$$

The advantage of basing the mean Nusselt number on the logarithmic mean temperature difference is that $Nu'(z)$ tends to the same limiting value as the local coefficient $Nu(z)$. For as $z \to \infty$, $\theta_M \sim \mathfrak{B}_1 \exp(-\lambda_1 z)$ and hence

$$Nu'(z) \sim \frac{1}{4}\left\{\lambda_1 + \frac{1}{z}\ln\left(\frac{1}{\mathfrak{B}_1}\right)\right\}. \tag{20}$$

The foregoing results are based on fundamental definitions given by Jakob [4] and are true for a duct of any cross-section.

Basis of the method of solution. We consider the general domain D'. The following is similar in principle to the method of Galerkin [5]. Let $\{\phi_n\}$ be the complete set of eigenfunctions of the equation

$$\nabla_1^2\phi + \Lambda\phi = 0, \tag{21}$$

with $\partial\phi/\partial\nu' = -N\phi$ on C'. Any arbitrary function $\Theta(x, y)$ which satisfies these boundary conditions and which possesses continuous partial derivatives up to the second order can be expanded in an absolutely and uniformly convergent series in the form

$$\Theta(x, y) = \sum_{m=0}^{\infty} a_m\phi_m(x, y), \tag{22}$$

where

$$a_m = \frac{1}{\delta_m(m)}\iint_{D'} \Theta\phi_m\, dx\, dy \tag{23}$$

and

$$\delta_p(m) = \iint_{D'} \phi_m\phi_p\, dx\, dy, \tag{24}$$

so that $\delta_p(m) = 0$ unless $m = p$. We can make Θ the solution of Eq. (9) so that multiplying this equation by ϕ_m and integrating over D' we have, by Eqs. (21) and (23),

$$\delta_m(m)\Lambda_m a_m = \lambda\iint_{D'} r(x, y)\phi_m\Theta\, dx\, dy, \qquad (m = 0, 1, 2, \cdots), \tag{25}$$

where $r(x, y) = w(x, y)/w_0$. If we substitute for Θ by Eq. (22) then Eq. (25) is reduced to an infinite set of homogeneous linear algebraic equations

$$\sum_{p=0}^{\infty}\{\delta_p(m)\Lambda_m - \lambda b_p(m)\}a_p = 0, \qquad (m = 0, 1, 2, \cdots), \tag{26}$$

where

$$b_p(m) = \iint_{D'} r\phi_m\phi_p\, dx\, dy. \tag{27}$$

The matrix associated with Eqs. (26) is symmetrical since $b_p(m) \equiv b_m(p)$ and the eliminant for a non-trivial solution gives an infinite determinantal equation $\Delta(\lambda) = 0$ whose latent roots are the eigenvalues of Eq. (9). Dividing each row of $\Delta(\lambda)$ by its leading diagonal element, the resulting determinant converges [6] if the off-diagonal sum is absolutely convergent and λ has no value which makes a leading diagonal element zero. If this condition is satisfied the convergence of $\sum_{p=0}^{\infty} \mid a_p \mid$ and, moreover,

of $\sum_{p=0}^{\infty} \delta_p(p)\ \Lambda_p \mid a_p \mid$ follows. The eigenvectors $\{a_p^{(n)}\}$ corresponding to a given root $\lambda = \lambda_n$ can then be obtained, theoretically, in terms of any arbitrary coefficient but in practice the determination of a given eignensolution is a problem in numerical analysis. One special point concerning the above formulae may be noted. It will be necessary in the rectangular case to specify solutions of Eq. (21) by number pairs, that is $\{\phi_{m,n}\}$ rather than $\{\phi_m\}$, and the expansion (22) is now a sum over all number pairs from $m, n = 0, 1, 2, \cdots$. Thus $b_p(m)$ in Eq. (26) is then written $b_{p,q}(m, n)$ and is associated with a coefficient $a_{p,q}$ in a double sum over number pairs from $p, q = 0, 1, 2, \cdots$. The equations hold for $m, n = 0, 1, 2, \cdots$, and $\delta_{p,q}(m, n)$, written for $\delta_p(m)$ in Eq. (24), is non-zero only if both $m = p$ and $n = q$.

The rectangular cross-section. In this case the boundary conditions become

$$\frac{\partial \Theta}{\partial x} = \pm N\Theta \quad \text{when} \quad x = \begin{matrix} 0 \\ l \end{matrix}, \qquad \frac{\partial \Theta}{\partial y} = \pm N\Theta \quad \text{when} \quad y = \begin{matrix} 0 \\ l' \end{matrix}, \tag{28}$$

where $l = (1 + \alpha)/2$, $l' = (1 + 1/\alpha)/2$ and α is the aspect ratio. Now the functions $X_m(x) \equiv \sin(\beta_m x + \beta_m')$, where $\tan \beta_m' = \beta_m/N$, $(0 \leq \beta_m' \leq \pi/2)$, satisfy the first of Eqs. (28) if $\beta_m (m = 0, 1, 2, \cdots)$, is a positive root (the negative roots only repeat the functions) of the equation

$$\tan \beta l = 2N\beta/(\beta^2 - N^2). \tag{29}$$

The roots of Eq. (29) form two separate sets which satisfy respectively the equations

$$\begin{aligned} \beta \tan \tfrac{1}{2}\beta l &= N \\ \text{and} \qquad N \tan \tfrac{1}{2}\beta l + \beta &= 0, \end{aligned} \tag{30}$$

the corresponding functions being respectively symmetrical and anti-symmetrical about $x = l/2$. The root $\beta = 0$ of the second equation does not contribute a function $X_m(x)$ but the root $\beta = 0$ of the first in the case $N = 0$ contributes a function $X_m(x) = 1$, which must be included for completeness. A similar set of functions $Y_n(y) \equiv \sin(\gamma_n y + \gamma_n')$, where $\tan \gamma_n' = \gamma_n/N$, $(0 \leq \gamma_n' \leq \pi/2)$, satisfy the second of Eqs. (28) if γ_n $(n = 0, 1, 2, \cdots)$, is a root of Eq. (29) with l' for l. Adopting the double suffix notation defined above, $\phi_{m,n} = X_m(x)\ Y_n(y)$ satisfies Eq. (21) with the results

$$\Lambda_{m,n} = \beta_m^2 + \gamma_n^2 \tag{31}$$

and

$$\delta_{m,n}(m, n) = \tfrac{1}{4}ll'\left(\frac{\beta_m^2 + N^2 + 2N/l}{\beta_m^2 + N^2}\right)\left(\frac{\gamma_n^2 + N^2 + 2N/l'}{\gamma_n^2 + N^2}\right). \tag{32}$$

We can now obtain a formula for $b_{p,q}(m, n)$, given by Eq. (27), in the form

$$b_{p,q}(m, n) = \tfrac{1}{4}\{c_{|m-p|,|n-q|} - c_{|m-p|,n+q} + c_{m+p,n+q} - c_{m+p,|n-q|}\}, \tag{33}$$

where

$$\begin{aligned} c_{m+p,n+q} = \int_0^l \int_0^{l'} r \cos\{(\beta_m + \beta_p)x \\ + (\beta_m' + \beta_p')\} \cos\{(\gamma_n + \gamma_q)y + (\gamma_n' + \gamma_q')\}\, dx\, dy, \end{aligned} \tag{34}$$

and a suffix $\mid m - p \mid$, say, involves a change of sign between elements with suffixes m and p on the right hand side of Eq. (34). Note that this notation is defined only for

compound suffixes. A term $c_{m,n}$ has no meaning in the general case, although exceptionally it has in the following limiting cases. Putting $N = \infty$, the constant wall temperature case, we have $\beta'_m \equiv 0$, $\gamma'_n \equiv 0$, $\beta_m = m\pi/l$, $\gamma_n = n\pi/l'$ and $\delta_{m,n}(m, n)\ \Lambda_{m,n} = (\pi^2/4\alpha)(m^2 + \alpha^2 n^2)$. The expansion (22) is now in the form of a double Fourier sine series in $(0 \leq x \leq l, 0 \leq y \leq l')$. The coefficients given by Eq. (34) can now be identified with members of the set whose general term is

$$d_{i,j} = \int_0^l \int_0^{l'} r \cos(i\pi x/l) \cos(j\pi y/l')\, dx\, dy, \qquad (i, j = 0, 1, 2 \cdots), \tag{35}$$

that is, they may be associated with the coefficients of the double Fourier cosine expansion of $r(x, y)$ in $(0 \leq x \leq l, 0 \leq y \leq l')$. Equation (33) still holds identically with d for c. On the other hand, if $N = 0$ the only difference is that $\beta'_m \equiv \gamma'_n \equiv \pi/2$ and we find that we can write d in place of c in Eq. (33) provided we change the negative signs in this equation. This case is of no interest in the present problem but may be so in other applications since the above formulae are true for arbitrary $r(x, y)$. In the present problem $r(x, y)$ is found from Eq. (3) to be

$$r(x, y) = \frac{1}{f_0} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} i^{-1} j^{-1} (i^2 + \alpha^2 j^2)^{-1} \sin(i\pi x/l) \sin(j\pi y/l'), \tag{36}$$

where

$$f_0 = (4/\pi^2) \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} i^{-2} j^{-2} (i^2 + \alpha^2 j^2)^{-1}$$

and, because $r(x, y)$ is symmetrical about both $x = l/2$ and $y = l'/2$, i and j are restricted to be odd integers only. Substitution into Eq. (34) yields a formula for $c_{m+p,n+q}$ which is expressed as a double summation with respect to i and j but which can be summed with respect to one of these variables of summation to give a rapidly convergent single series. We also find that, because of the symmetry present in $r(x, y)$, $c_{m+p,n+q}$ is zero under certain circumstances. Let us associate odd integers with the roots of the first of Eqs. (30) and even integers with the roots of the second. Then it is readily shown that $c_{m+p,n+q}$ is non-zero only if $m + p$ and $n + q$ are both even integers, and that the same applies to each of the other three coefficients in Eq. (33). The formula for the non-zero coefficients is

$$c_{m+p,n+q} = c'_{m+p,n+q} \sum_{j=1}^{\infty} \frac{\pi\alpha j \tan\{\frac{1}{2}(\beta_m + \beta_p)l\} - (\beta_m + \beta_p)l \tanh(\frac{1}{2}\alpha j\pi)}{j\{\alpha^2 j^2 \pi^2 + l^2(\beta_m + \beta_p)^2\}\{\pi^2 j^2 - l'^2(\gamma_n + \gamma_q)^2\}}, \tag{37}$$

where $c'_{m+p,n+q} = \pi^3 l' \cos(\beta'_m + \beta'_p) \cos(\gamma'_n + \gamma'_q)/\alpha f_0 (\beta_m + \beta_p)$, and j is odd. The other three coefficients required in Eq. (33) are obtained from Eq. (37) by appropriate changes of sign. Since, for all values of N, the roots of Eqs. (30) all approach values which are integral multiples of π/l, it is clear that the quadruple sum of $b_{p,q}(m, n)$ with respect to p, q, m, and n converges absolutely so that $\Delta(\lambda)$ converges.

We may now consider the special nature of the eigenfunctions derived from the solutions of the algebraic equations. Since $b_{p,q}(m, n)$, which is the coefficient of $a_{p,q}$ in the (m, n)th equation, is non-zero only if $m + p$ and $n + q$ are both even it follows at once that the equations break up into four independent sets. Since also $a_{p,q}$ is the coefficient of $X_p(x)Y_q(y)$ in the expansion for Θ there are four corresponding independent sets of eigenfunctions which exhibit the alternative properties (*i*) symmetry about both

$x = l/2$ and $y = l'/2$, (ii) anti-symmetry about $x = l/2$ and $y = l'/2$, (iii) symmetry about $x = l/2$ with anti-symmetry about $y = l'/2$ and (iv) the opposite of the last case. Of these only (i) concerns us in this problem since by Eq. (13) only these solutions give non-zero coefficients in Eq. (12). The special case $\alpha = 1$ needs further consideration since here, in addition to the matrix symmetry $b_{p,q}(m, n) = b_{m,n}(p, q)$ present in all cases, we also have $b_{p,q}(m, n) = b_{q,p}(n, m)$. It follows that if an ordered set of coefficients $\{a_{p,q}\}$ with a given eigenvalue $\lambda = \lambda_n$ satisfies the algebraic equations then so, with the same eigenvalue, does the set $\{a_{q,p}\}$ obtained by interchange of $a_{q,p}$ with $a_{p,q}$. That is, if $\Theta'_n(x, y) = \sum_{p=1}^{\infty} \sum_{q=1}^{\infty} a_{p,q} X_p(x) Y_q(y)$ is a solution then a linearly independent solution is $\Theta'_n(y, x) = \sum_{p=1}^{\infty} \sum_{q=1}^{\infty} a_{q,p} X_p(x) Y_q(y)$ and, since the eigenvalues are equal, these solutions may not satisfy the orthogonality property given by Eq. (10), which would invalidate Eq. (13). On the other hand the sum and difference of these solutions are both themselves solutions and we can write their contribution to the right hand side of Eq. (12) as

$$\mathfrak{A}_n\{\Theta'_n(x, y) + \Theta'_n(y, x)\} + \mathfrak{A}'_n\{\Theta'_n(x, y) - \Theta'_n(y, x)\}.$$

Multiplying Eq. (12) by $w\{\Theta'_n(x, y) - \Theta'_n(y, x)\}$ and integrating over D' we find at once that $\mathfrak{A}'_n = 0$ and in the remaining term, considered as a single eigenfunction with eigenvalue λ_n, the terms $X_p(x)Y_q(y), X_q(x)Y_p(y)$ occur with equal coefficients. It follows that in the case $\alpha = 1$ we can *ab initio* put $a_{p,q} = a_{q,p} (p, q = 1, 3, 5, \cdots)$, in the algebraic equations and that each eigenfunction derived from this reduced set of equations corresponds to a unique term in the expansion (12) with $\mathfrak{A}_n$ given as usual by Eq. (13). We have, of course, assumed that the λ_n of the reduced set of equations are themselves distinct. It follows also in this case that the expansion (8) consists only of functions for which $\Theta(x, y) \equiv \Theta(y, x)$, which we would expect physically.

It remains only, in the general case, to evaluate $\mathfrak{A}_n$ from each computed Θ_n. Substituting in Eq. (13) from Eq. (17) we have

$$\mathfrak{A}_n = -\frac{1}{A'\lambda_n} \int_{C'} \frac{\partial \Theta_n}{\partial \nu'} ds'$$

and since

$$\int_{C'} \frac{\partial \Theta_n}{\partial \nu'} ds' = -2\left[\int_0^{l} \left\{\frac{\partial \Theta_n}{\partial y}\right\}_{y=0} dx + \int_0^{l'} \left\{\frac{\partial \Theta_n}{\partial x}\right\}_{x=0} dy\right] \tag{38}$$

then

$$\mathfrak{A}_n = \frac{16\alpha N^2}{(1+\alpha)^2\lambda_n} \sum_{p=1}^{\infty} \sum_{q=1}^{\infty} \sin \beta'_p \sin \gamma'_q (\beta_p^{-2} + \gamma_q^{-2}) a_{p,q}^{(n)}. \tag{39}$$

A more rapidly convergent formula is found by substituting directly for Θ_n into Eq. (13) but it is more complicated except in the special case $N = \infty$, in which it becomes

$$\mathfrak{A}_n = \frac{1}{4f_0} \sum_{p=1}^{\infty} \sum_{q=1}^{\infty} p^{-1} q^{-1} (p^2 + \alpha^2 q^2)^{-1} a_{p,q}^{(n)}. \tag{40}$$

In these formulae $\{a^{(n)}_{p,q}\}$ are the particular set of coefficients which refer to Θ_n and which satisfy Eq. (11). In practice, solutions of the algebraic equations have been computed by arbitrarily putting one coefficient equal to unity. If $\{A^{(n)}_{p,q}\}$ is such a solution and we put $A_{p,q}^{(n)} = \mathfrak{N}_n a_{p,q}^{(n)}$, then $\mathfrak{N}_n$ is found from Eq. (11). From Eq. (9) we

have $\int\int_D w\Theta_n^2\,dx\,dy = -(w_0/\lambda_n)\int\int_D \Theta_n\,\nabla^2\Theta_n\,dx\,dy$ so that

$$\mathfrak{N}_n^2 = \frac{4\alpha}{\lambda_n(1+\alpha)^2}\sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\delta_{p,q}(p,q)(\beta_p^2+\gamma_q^2)[A_{p,q}^{(n)}]^2. \tag{41}$$

Computational results. For a given value of α the computation from the algebraic equations of the first few eigensolutions, arranged in ascending order of λ, is a standard problem. We have used relaxation methods in conjunction with Rayleigh's principle. λ_n for a given eigensolution being estimated by the Rayleigh quotient

$$\begin{aligned}\lambda_n &= -\int_0^{l}\int_0^{l'}\Theta_n\nabla_1^2\Theta_n\,dx\,dy\Big/\int_0^{l}\int_0^{l'}r\Theta_n^2\,dx\,dy,\\ &= \sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\delta_{p,q}(p,q)(\beta_p^2+\gamma_q^2)[A_{p,q}^{(n)}]^2\Big/\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sum_{p=1}^{\infty}\sum_{q=1}^{\infty}b_{p,q}(i,j)A_{i,j}^{(n)}A_{p,q}^{(n)}.\end{aligned} \tag{42}$$

Since the diagonal elements strongly dominate the matrix, good initial approximations to the eigenvalues are found by equating each successive diagonal to zero giving estimates $\lambda_{m,n} = \delta_{m,n}(m,n)(\beta_m^2+\gamma_n^2)/b_{m,n}(m,n)$ where, for obvious reasons, we temporarily adopt double-suffix notation for the λ's. We obtain an initial estimate of the (m,n)th eigenfunction using this approximation for $\lambda_{m,n}$, arbitrarily putting $A_{m,n}^{(m,n)} = 1$ and using the other equations except the (m,n)th to estimate the remaining coefficients in the form

$$A_{p,q}^{(m,n)} \doteq b_{m,n}(p,q)\Big/\left[\frac{\delta_{p,q}(p,q)(\beta_p^2+\gamma_q^2)}{\delta_{m,n}(m,n)(\beta_m^2+\gamma_n^2)}\,b_{m,n}(m,n) - b_{p,q}(p,q)\right]. \tag{43}$$

This approximation is computed until the coefficients attain some prescribed order of smallness acceptable as error. It is then improved by the relaxation process until a final accepted solution is reached upon which we impose the criteria that no coefficient larger than the agreed order of smallness be omitted, nor must any omitted coefficient affect, to the agreed working accuracy, the value of any included coefficient. The eigenfunctions themselves are of little physical interest but to illustrate the numerical analysis we give in Table 1, correct to four decimal places, the first two in the case $\alpha = 1$, $N = \infty$. The physical results of interest depend on the expansions (14), (18) and (19). The $\mathfrak{G}$'s converge fairly rapidly and we need only the first few terms to describe the physical domain, that is, $z > z_0$ (say). It is pointless to extend the solution too near to $z = 0$ for since $z = \zeta/d\text{Pé}$

TABLE 1

$\alpha = 1$, $N = \infty$

p \ q	1	3	5	7	9					
1	1.0000	−0.0369	−0.0022	−0.0004	−0.0001				−0.0001	17
3	−0.0369	0.0043	0.0003						−0.0002	15
5	−0.0022	0.0003							−0.0003	13
7	−0.0004			Θ_1					−0.0011	11
9	−0.0001				Θ_2			0.0001	0.0001	−0.0021	9
						0.0001	0.0004	0.0001	−0.0068	7
					0.0001	0.0004	−0.0002	0.0334	−0.1583	5
					0.0001	0.0001	0.0334	−0.2967	1.0000	3
	−0.0001	−0.0002	−0.0003	−0.0011	−0.0021	−0.0068	−0.1583	1.0000	0.3711	1
	17	15	13	11	9	7	5	3	1	q \ p

then, for fixed ζ and a given fluid, this would be equivalent to increasing the Reynolds number beyond the laminar flow range. In our results we have treated the constant wall temperature case in considerable detail since this is the case dealt with by Clark and Kays (*loc. cit.*). In Table 2 the first three λ_n and associated $\mathcal{B}_n$ are given in each of the cases $\alpha = 1, 2/3, 1/2, 1/4$, and $1/8$. For the more general radiation boundary condition we have considered only the square duct. The first three terms for θ_M in the case $N = 2$ are

$$\theta_M = 0.972 \exp(-4.81z) + 0.023 \exp(-47.6z) + 0.003 \exp(-127z) \cdot + \cdots$$

while the first only for the cases $N = 10, 20$ are respectively

$$\theta_M = 0.893 \exp(-9.18z) + \cdots \quad \text{and} \quad \theta_M = 0.860 \exp(-10.37z) + \cdots$$

The eigenvalues are well separated for the square duct so that in fact even the first term describes well the physical domain. Beyond $N = 20$, λ_1 [and hence the important quantity $Nu(\infty)$] can be calculated to good accuracy from the formula

$$\frac{(\lambda_1)_N}{(\lambda_1)_{N=\infty}} \doteq \frac{\beta_1^2}{\pi^2}\left(\frac{\beta_1^2 + N^2 + 2N}{\beta_1^2 + N^2}\right)^2 \frac{[b_{1,1}(1,1)]_{N=\infty}}{[b_{1,1}(1,1)]_N}. \tag{44}$$

This formula is based on the assumption that the correct value of λ_1 for a given N bears the same ratio to λ_1 in the case $N = \infty$ as do the corresponding diagonal estimates of these eigenvalues in the initial approximation given previously since, by Rayleigh's principle, these latter are always over-estimates. For example when $N = 20$, Eq. (44) gives $\lambda_1 = 10.42$ against the correct value $\lambda_1 = 10.37$.

Comparison of thermal results. The theory used by Clark and Kays is based on the assumption, previously used by Seban and Shimazaki [7] for turbulent flow in cylinders, that far enough from the thermal inlet

$$\frac{\partial}{\partial \zeta}\left(\frac{T - T_1}{T_M - T_1}\right) = 0$$

and this is borne out in our work since, for large enough z, $\theta/\theta_M \sim \Theta_1(x, y)/\mathcal{C}_1$, and is independent of z. Calculated results for the limiting Nusselt number in the cases $\alpha = 1$ and 0.5 are given respectively as $Nu'(\infty) = 2.89$ and 3.39 and these compare well with our values of 2.98 and 3.39. No theoretical information on the variation of Nusselt number in the thermal inlet length is given, but experimental data have been obtained by the authors for the aspect ratios $\alpha = 1$ and 0.382. These largely confirm the theoretical

TABLE 2

$N = \infty$

α	1.000	0.667	0.500	0.250	0.125
λ_1	11.91	12.49	13.57	17.76	22.38
λ_2	71.07	51.58	41.17	28.17	25.61
λ_3	157.9	99.71	94.93	47.82	31.81
$\mathcal{B}_1$	0.804	0.802	0.789	0.756	0.737
$\mathcal{B}_2$	0.104	0.064	0.071	0.107	0.091
$\mathcal{B}_3$	0.014	0.043	0.020	0.028	0.034

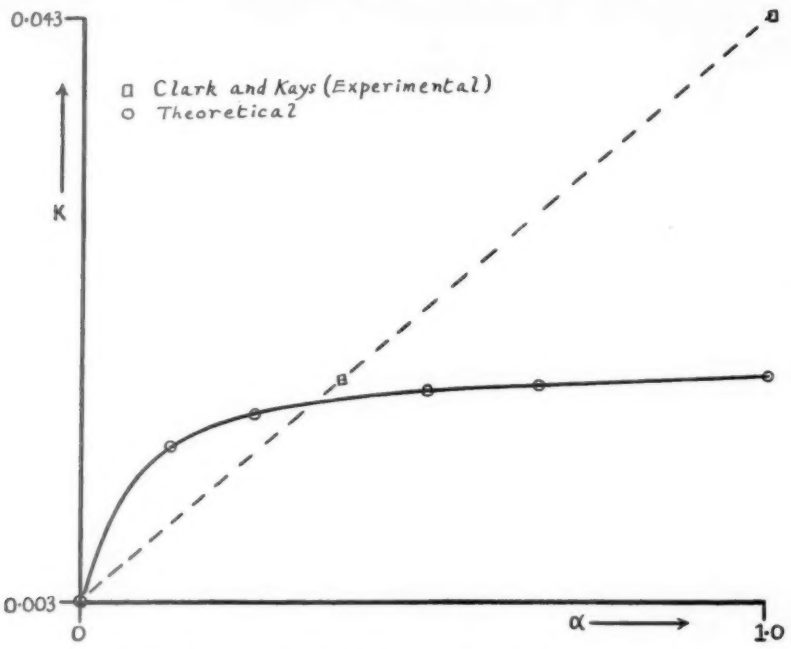


FIG. 1. K against aspect ratio for the case of constant wall temperature.

values of $Nu(\infty)$ and also it is found that for small values of $d\text{Pé}/\zeta$ the variation of the logarithmic mean Nusselt number is linear according to the formula

$$\frac{Nu'(z)}{Nu'(\infty)} = 1 + K\left(\frac{d\,\text{Pe}'}{\zeta}\right). \tag{45}$$

This linear law follows theoretically from Eq. (20) which gives the theoretical value $K = (1/\lambda_1)\ ln\ (1/\mathfrak{B}_1)$. In the case $\alpha = 0.382$ the theoretical value $K = 0.016$ agrees well with the experimental value 0.017, but this is not so for $\alpha = 1$ where we find $K = 0.018$ against the experimental value 0.042. The theoretical curve for K against α is compared with Clark and Kays tentative linear correlation in Fig. 1. The disagreement is serious, but we must point out that the experimental curve is determined by only two observed points, the result for $\alpha = 0$ being theoretical, so that an error in an observed point could give a very different curve. On account of the discrepancy for the square duct we have investigated the radiation boundary condition fully in this case and the results for K against N^{-1} are given in Fig. 2. Clearly the values of K are always lower than those in the constant wall temperature case, so no possible explanation is forthcoming from these results. On the other hand, comparison of experimental and theoretical values of K for the circular cross-section [8] suggests that experimental values may be considerably higher. This may lessen the discrepancy in the square case but, for consistency, the experimental value for $\alpha = 0.382$ should also be higher. This is possible since Clark and Kays state that in this case the ratio of duct-length to mean hydraulic depth used in their apparatus could possibly be higher than the assumed value by as much as 100%;

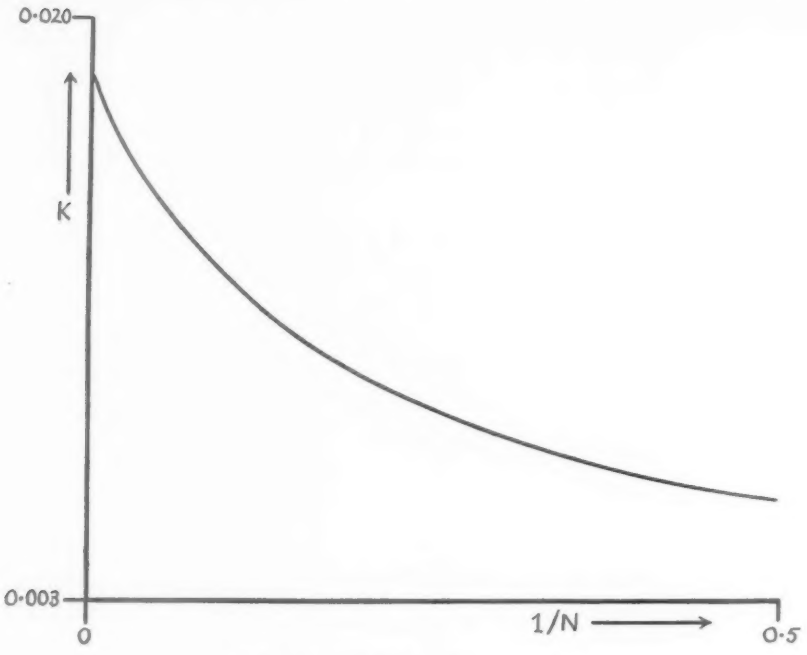


FIG. 2. K against N^{-1} for the square duct.

this would lead to a larger value of K. In the absence of more detailed experimental results, however, it is impossible to state precisely the cause of the disagreement. Finally, we should notice that there is some doubt regarding the theoretical value for K near the limiting case $\alpha = 0$. We consider only $N = \infty$ but the general case is similar. If we keep the side of the duct parallel to the ξ-axis fixed and let the other become large then $\partial^2\theta/\partial y^2 \to 0$, $w/w_0 \to 24\, x(1/2 - x)$ in Eq. (4). The solution for θ may now be written $\theta(x, z) = \sum_{m=1}^{\infty} \mathcal{A}_m' \vartheta_m(x) \exp(-\lambda_m' z)$ where $\vartheta'' + 24\, \lambda' x(1/2 - x)\, \vartheta = 0$ with $\vartheta(0) = \vartheta(1/2) = 0$. Now each $\vartheta_m(x)$ can be written as

$$\frac{4}{\pi} \sum_{n=1}^{\infty} \vartheta_m \frac{\sin(n\pi y/1')}{n}, \qquad (n = 1, 3, 5 \cdots),$$

that is, it can be considered as a sum of functions $\Theta_{m,n}(x, y)$ with identical eigenvalues λ_m' and these latter functions can, for varying m and n, be identified with the limiting solutions, here written in double-suffix notation, of our previous algebraic equations. It is therefore clear that when $\alpha = 0$ the value of $\mathcal{B}_1$ to be used in Eq. (20) should be the sum $\sum_{n=1}^{\infty} \mathcal{B}_{1,n}$, $(n = 1, 3, 5, \cdots)$, of the $\mathcal{B}$'s associated with $\Theta_{1,n}(x, y)$ and since it is easily shown that $\mathcal{B}_{1,n} = n^{-2}\, \mathcal{B}_{1,1}$ this sum is $\pi^2\, \mathcal{B}_{1,1}/8$. When $\alpha \neq 0$ and the λ's are all distinct we should use $\mathcal{B}_{1,1}$ which, in double-suffix notation, corresponds to the smallest λ. There is therefore some doubt as to the correct procedure near $\alpha = 0$ but, in practice, it can make very little difference to Fig. 1 since K is so small at this end of the curve.

Acknowledgment. We acknowledge a grant in aid of this work by the Royal Society. A preliminary (unpublished) account was given by one of us (A.Mc.D.M.) to the IXth International Congress of Mechanics and Applied Mathematics, Brussels, 1956.

References

1. S. H. Clark and W. M. Kays, Trans. A.S.M.E. **75**, 859–866 (1953)
2. W. M. Hampton, Nature **157**, 481 (1946)
3. R. Courant and D. Hilbert, *Methods of mathematical physics*, vol. 1, Interscience publishers, New York, 1953
4. M. Jakob, *Heat transfer*, vol. 1, John Wiley, New York, 1949
5. W. E. Milne, *Numerical solution of differential equations*, John Wiley, New York, 1953, p. 114
6. E. T. Whittaker and G. N. Watson, *Modern analysis*, Cambridge, 1927, pp. 36, 37, 417
7. R. A. Seban and T. T. Shimazaki, Trans. A.S.M.E. **73**, 803–809 (1951)
8. W. M. Kays and A. L. London, Trans. A. S. M. E. **74**, 1179–1189 (1952)

BOOK REVIEWS

(*Continued from p. 284*)

estimated first mode frequencies [by the Raleigh method] will ordinarily be too high". One could add many more examples to this list of incorrect or misleading statements.

In some respect, the book lacks order. Aside from questions of sequence of subject matter which appears to the reviewer often unmotivated, one finds a portion of a discussion of the nonlinear pendulum problem, as well as that of cases of nonlinear damping, in Chapter 2 and 4, while nonlinear vibrations are discussed in Chapter 7. In Chapter 12, dealing with vibrating beams and related subjects, one finds that "mention of damped vibration with many degrees of freedom has been avoided up to this point . . ." while the dynamic vibration absorber with damping, and the Lanchester damper are treated in Chapter 9.

Perhaps more serious than the errors of commission are those of omission. One of these is the virtual absence of references. Instead, each chapter is followed by a list of "suggested reading". These lists show a serious disregard, both for important work done in the field, and for the relation of the elementary level of the book under review to the advanced level of some of the suggested readings. As one example for each may serve (1) the omission of McLachlan's book on "Ordinary non-linear differential equations" in the section on nonlinear vibrations, and (2) the inclusion of Mindlin's paper on the "Influence of rotary inertia and shear on flexural motions of isotropic elastic plates". This latter is suggested reading for students from whom even the elementary equation of motion of the one-dimensional vibrating beam was withheld in the body of the book, and who have been left completely in the dark on the existence of rotary inertia and transverse shear effects in beams.

No mention is made of Lagrange's equations of motion or of generalized coordinates; in fact, there is no treatment of theoretical mechanics. The section on nonlinear vibrations contains no phase-plane considerations, no perturbation method, no iteration method and, not a single differential equation. Duffing's name or equation are not mentioned and, although there is a paragraph on relaxation oscillations, Van der Pol's name or equation are also overlooked. The section on vibrating beams, plates and membranes (in that order) fails to derive or even to state the equations of motion of any of these systems. In the treatment of plates, Rayleigh's method is applied to an approximate energy expression without stated restrictions on plate thickness or vibrational amplitudes. In a discussion of the limitations of beam theory as given in the book, reference is made to "massive extensions . . . attached to the beam", and to gyroscopic action, but none to rotary inertia or transverse shear. A chapter on control systems contains neither operational methods nor transfer functions, and the suggested reading includes "Weiner, Norbert, Cybernetics". A final chapter on fatigue presents some largely descriptive paragraphs; phenomenological or solid state physics theories and statistical approaches are ignored.

R. M. Rosenberg

Handbook of supersonic aerodynamics. Volumes 1, 2, 3, 4, 5. Supt. of Documents, U. S. Government Printing Office, Washington 25, D. C.

The purpose of this Handbook is to provide a compilation of formulae, tables and other related information which would be of use to 'designers of supersonic vehicles.' The volumes are issued in loose-leaf form and all are not as yet complete, provision having been made for the insertion of additional material as it becomes available. When completed, this series will consist of six volumes containing 21 sections. Some idea of their content may be gathered from the following list of titles of sections which have thus far appeared:

Vol. 1: Section 1—Symbols and nomenclature; Section 2—Fundamental equations and formulae; Section 3—General atmospheric data; Section 4—The mechanics and thermodynamics of steady one-dimensional gas flow. Vol. 2: Section 5—Compressible flow tables and graphs. Vol. 3: Section 6—Two-dimensional airfoils. Vol. 4: Section 12—Aeroelastic phenomena. Vol. 5: Section 15—Properties of gases.

W. H. Reid

(*Continued on p. 321*)

THE EXACT SOLUTION OF BORDA'S MOUTHPIECE IN TWO DIMENSIONS*

BY

CHARLES A. HACHEMEISTER

(*Polytechnic Institute of Brooklyn*)

The solution of the semi-infinite two-dimensional Borda mouthpiece is well known. Its coefficient of contraction σ is 0.5. The solution of the finite mouthpiece seems to have been avoided, apparently because it involves elliptic integrals.

The finite Borda mouthpiece, Fig. 1, is formed by two walls DC and $D'C$ of length b

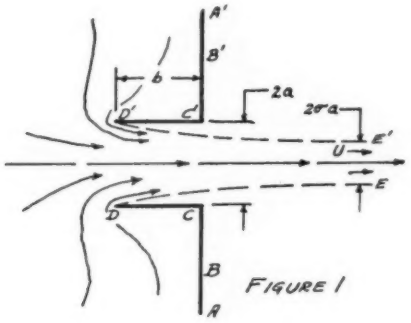


FIGURE 1

projecting into a semi-infinite reservoir which is bounded by the semi-infinite walls $A'B'C'$ and ABC having a gap of width $2a$ between them. Inviscid incompressible fluid flows out of the reservoir through the mouthpiece to form a jet which is bounded by the free stream lines $D'E'$ and DE. The jet contracts to the width $2\sigma a$ at $E'E$ far from the mouthpiece where the speed of the fluid is uniform and of value U. The total efflux from the reservoir is therefore $2\sigma aU$. The speed of the fluid along the free stream lines $D'E'$ and DE is U. Along the wet side of the mouthpiece and reservoir walls, $A'B'C'D'$ and $ABCD$, the surface speed varies. It is zero at the corners C' and C and at the infinite points A' and A. At two places, B' and B, the surface speed has a maximum. This follows because B is between A and C where the surface speed is zero.

The values of the coefficient of contraction σ and the maximum surface speed with its location are of interest.

Because the flow is in the general category of "potential flow", the techniques employing conformal mapping are applicable. This principle of analysis is not new.

Starting with the lower half of the symmetrical geometry of Fig. 1 redrawn on the z-plane of Fig. 2a (the mapping figures are grouped together on a later page) define: the complex potential $P = \phi + \psi i$, where ϕ is the potential function, and ψ is the stream function, so that the z-plane velocity $V_z = V_x + V_y i = -(dP/dz)^*$ where the * means conjugate.

*Received August 12, 1958.

A pair of mapping sequences, one set involving the potential P, the other the potential derivative dP/dz, is formed to coalesce finally, establishing the desired formulas. This procedure is indicated by the diagram, Figs. 2*a* through 2*e*, indicating the sequential transformations.

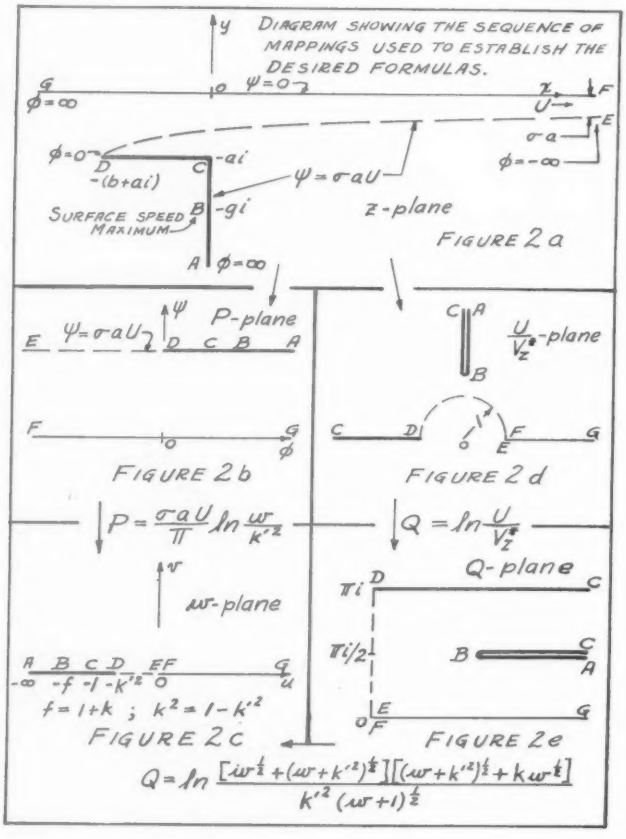
The potential sequence is started by assigning to the potential function ϕ and to the stream function ψ values which are consistent with the existing conditions at specific places in the z-plane:

Along the left infinite semi-circle $\phi = \infty$ (including points A and G), at the point D $\phi = 0$, at points E, F $\phi = -\infty$, along the stream line $ABCDE$ $\psi = \sigma aU$, along the stream line FG $\psi = 0$.

These values are then plotted to form the P-plane map of Fig. 2*b*. The P-plane strip is then mapped onto the upper half of the w-plane of Fig. 2*c* by means of the Schwarz-Christoffel transformation

$$dP/dw = \sigma aU/\pi w; \qquad P = (\sigma aU/\pi) \ln (w/k'^2). \tag{1}$$

The choice of -1 and $-k'^2$ as the w-plane locations of C and D respectively is made in the light of subsequent developments.



The potential derivative (velocity) sequence of mappings is started by utilizing certain known facts about the velocity in the z-plane. Rather than mapping the potential derivative, the reciprocal of the conjugate velocity normalized with respect to the terminal velocity U of the jet is mapped. This function maps into a simple geometry. Because

$$V_z^* = -dP/dz = q \exp(-\theta i)$$

where q is the speed and θ is the direction of the velocity,

$$U/V_z^* = (U/q) \exp(\theta i).$$

Along the rigid walls and the stream line of symmetry the direction, θ, is dictated by these boundaries. Along the free stream line the speed q is constant. The values of U/V_z^* along the boundaries can then be tabulated:

Along	ABC	CD	DE	FG
$U/V_Z^* =$	$(U/q)\exp(\pi i/2)$	$(U/q)\exp(\pi i)$	$\exp(\theta i)$	U/q
with	$\infty > U/q > U/q_B$	$\infty > U/q > 1$	$\pi > \theta > 0$	$\infty > U/q > 1$

These values are then plotted to form the U/V_z^*-plane map of Fig. 2d.

The upper half of the U/V_z^*-plane excluding the unit semi-circle is then mapped onto the semi-infinite strip in the Q-plane of Fig. 2e by means of the transformation

$$Q = \ln (U/V_z^*). \tag{2}$$

Finally the two mapping sequences are coalesced by mapping the Q-plane strip onto the upper half of the w-plane of Fig. 2c applying the Schwarz-Christoffel transformation

$$dQ/dw = N(w + f)/[w^{\frac{1}{2}}(w + k'^2)^{\frac{1}{2}}(w + 1)].$$

Both N and f are evaluated by integrating between G to A and C to C in the Q-plane and along the corresponding semi-circles in the w-plane. The results are:

$$N = \tfrac{1}{2} \qquad f = 1 + k \quad \text{with} \quad k^2 = 1 - k'^2.$$

Putting these into the expression for dQ/dw, integrating and making appropriate adjustment for correspondence of points, there results the Q to w-plane transformation

$$Q = \ln \frac{[w^{\frac{1}{2}} + (w + k'^2)^{\frac{1}{2}}][(w + k'^2)^{\frac{1}{2}} + kw^{\frac{1}{2}}]}{k'^2(w + 1)^{\frac{1}{2}}}. \tag{3}$$

The connection between the w- and z-plane is effected by substituting the potential derivative for the velocity in Eq. (2) and using the potential derivative from Eq. (1)

$$Q = \ln \frac{U}{V_z^*} = \ln \left(-U \frac{dz}{dP}\right) = \ln \left(-U \frac{dz}{dw}\frac{dw}{dP}\right) = \ln \left(-\frac{\pi w}{\sigma a}\frac{dz}{dw}\right). \tag{4}$$

Equating the arguments of the logarithms in Eqs. (3) and (4) establishes, after rearrangement, the z to w-plane transformation in derivative form

$$-\frac{\pi}{\sigma a}\,dz = \frac{[w^{\frac{1}{2}} + (w + k'^2)^{\frac{1}{2}}][(w + k'^2)^{\frac{1}{2}} + kw^{\frac{1}{2}}]}{k'^2 w(w + 1)^{\frac{1}{2}}}\,dw. \tag{5}$$

Equation (5) can be integrated to form the z to w-plane transformation. The result involves elliptic integrals of modulus k or k' with arguments that are complex, real or imaginary depending on the locations of the points in the z and w-plane. It is much easier to integrate Eq. (5) between limits which correspond to specific points, arranging the integrand in each case so that the resulting elliptic integrals have real arguments. Three such integrations suffice to establish the relationships among the parameter k, the mouthpiece length (b/a), the coefficient of contraction σ, and the location of the maximum surface speed (g/a).

The relation between the mouthpiece length (b/a) and the parameter k is obtained by integrating Eq. (5) between limits corresponding to the locations of points C and D in the z- and w-planes

$$-\frac{\pi}{\sigma a}\int_{-ai}^{-(b+ai)} dz = \frac{1+k}{k'^2}\int_{-1}^{-k'^2}\frac{w^{\frac{1}{2}}\,dw}{[(w+k'^2)(w+1)]^{\frac{1}{2}}} + (1+k)$$
$$\cdot\int_{-1}^{-k'^2}\frac{dw}{[w(w+k'^2)(w+1)]^{\frac{1}{2}}} + \frac{1+k}{k'^2}\int_{-1}^{-k'^2}\frac{dw}{(w+1)^{\frac{1}{2}}} + \int_{-1}^{-k'^2}\frac{dw}{w(w+1)^{\frac{1}{2}}}.$$

After performing the indicated integrations the results are arranged to

$$\frac{\pi b}{2\sigma a} = \frac{k}{1-k} + \frac{E(k) - k'^2K(k)}{1-k} - \tfrac{1}{2}\ln\frac{1+k}{1-k}. \tag{6}$$

Integrating between limits corresponding to the locations of points D and E establishes the relation between the coefficient of contraction σ and the parameter k. Here the infinite value of x is avoided by approaching point E in the z-plane:

$$\lim_{x\to\infty} -\frac{\pi}{\sigma a}\int_{-(b+ai)}^{(x-ai)} dz = \frac{1+k}{k'^2}\int_{-k'^2}^{0}\frac{(w+1)^{\frac{1}{2}}\,dw}{[w(w+k'^2)]^{\frac{1}{2}}} - \frac{k^2(1+k)}{k'^2}$$
$$\cdot\int_{-k'^2}^{0}\frac{dw}{[w(w+k'^2)(w+1)]^{\frac{1}{2}}} + \frac{1+k}{k'^2}\int_{-k'^2}^{0}\frac{dw}{(w+1)^{\frac{1}{2}}} + \int_{-k'^2}^{0}\frac{dw}{w(w+1)^{\frac{1}{2}}}.$$

The real parts are obviously infinite and so are ignored. The imaginary parts yield the desired formula:

$$\frac{\pi}{2\sigma} = (1+k)\frac{E(k') - k^2K(k')}{k'^2} + \frac{\pi}{2}. \tag{7}$$

Eliminating σ from Eqs. (6) and (7) establishes the relationship between (b/a) and the parameter k

$$\frac{b}{a} = \frac{k + E(k) - k'^2K(k) - [(1-k)/2]\ln[(1+k)/(1-k)]}{E(k') - k^2K(k') + \pi(1-k)/2}. \tag{8}$$

It is interesting to note that $k = 1$, $k' = 0$ corresponds to $b/a = \infty$, the semi-infinite Borda mouthpiece. Putting these values in Eq. (7) yields $\sigma = 1/2$, the well known coefficient for this mouthpiece. Similarly $k = 0$, $k' = 1$ corresponds to $b/a = 0$, the mouthpiece vanishes and discharge is through the remaining slit in the reservoir wall. Equation (7) now yields $\sigma = 1/(1 + 2/\pi)$, the known value for discharge through a slit.

The location of the maximum surface speed along the reservoir wall is found by integrating (5) between limits corresponding to the locations of points B and C

$$-\frac{\pi}{\sigma a}\int_{-g\,i}^{-a\,i} dz = \frac{-(1+k)i}{k'^2}\int_{-(1+k)}^{-1}\frac{(-w)^{\frac{1}{2}}\,dw}{[(-k'^2-w)(-1-w)]^{\frac{1}{2}}} + (1+k)i \cdot\int_{-(1+k)}^{-1}\frac{dw}{[-w(-k'^2-w)(-1-w)]^{\frac{1}{2}}} - \frac{(1+k)i}{k'^2}\cdot\int_{-(1+k)}^{-1}\frac{dw}{(-1-w)^{\frac{1}{2}}} - i\int_{-(1+k)}^{-1}\frac{dw}{w(-1-w)^{\frac{1}{2}}}$$

and

$$\frac{\pi(g-a)}{2\sigma a} = \frac{(1+k)[k^2K(k')-E(k')]}{k'^2} - \tan^{-1}k^{\frac{1}{2}} + \frac{1+k^{\frac{1}{2}}}{2(1-k^{\frac{1}{2}})}. \tag{9}$$

Equation (9) with (7) determines the value of g/a in terms of the parameter k and so the relative location of the maximum surface speed.

The value of the maximum surface speed is found using the potential derivative in terms of the transformation derivatives of Eqs. (1) and (5) evaluated for $w = -(1+k)$, the w-plane location of point B:

$$V_z^* = -\frac{dP}{dz} = -\frac{dP}{dw}\frac{dw}{dz}.$$

Substituting from Eqs. (1) and (5) and making $w = -(1+k)$

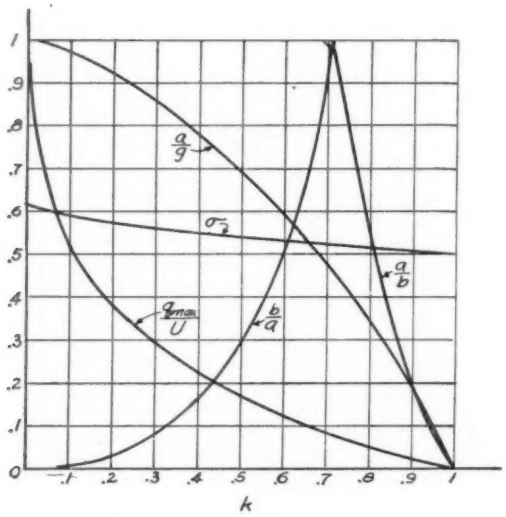


FIG. 3. Graphical representation of Equations (7), (8), (9), and (10) showing the relations of σ, the coefficient of contraction; b/a, the mouthpiece length; g/a, the location of the maximum surface speed; and q_{max}/U, the maximum surface speed, to the parameter k. $2a$ is the width of the mouthpiece. U is the terminal speed of the jet. The values on any ordinate correspond. See Figures 1 and 1a.

$$V_z = \frac{1 - k^4}{1 + k^4} Ui, \tag{10}$$

the velocity maximum at point B.

Eqs. (6), (7), (9) and (10) constitute the desired relationships. The results are shown graphically in Fig. 3.

For computation purposes it is convenient to introduce into Eqs. (6), (7) and (9)

$$B = \frac{E(k) - k'^2 K(k)}{k^2}$$

which is tabulated in *Tables of functions* by Jahnke and Emde, Dover Publications, 1945.

The following formulas from *Handbook of elliptic integrals for Engineers and Physicists*, Byrd and Friedman, Springer, 1945, were used for evaluating the various elliptic integrals:

Equation	Formula number
(6)	233.00, 233.01, 110.06, 1110.07
(7)	236.00, 236.01, 110.06, 110.07
(9)	232.06, 321.02, 232.00, 111.03, 122.10

References

1. L. M. Milne-Thomson, *Theoretical hydrodynamics*, Macmillan Co., 3rd ed. 1955
2. E. Weber, *Electro-magnetic fields*, Wiley 1950—Secs. 25-27, vol. 1

ON A FREE BOUNDARY VALUE PROBLEM FOR THE HEAT EQUATION

BY

WALTER T. KYNER*

University of Southern California

1. Introduction. W. L. Miranker [1] recently published an existence theorem for a free boundary value problem for the heat equation. Using a method due to I. Kolodner [2], he obtained a functional equation for the free boundary function $R(t)$ and showed that the existence of a solution to the functional equation implied the existence of the solution to the free boundary problem. He then solved the functional equation by an iterative method.

The mathematical problem which Miranker solved represents the heating of a long insulated metal rod which has begun to melt at one end ($x = 0$) and *after* a layer of liquid metal A units thick has formed, heat is applied at $x = 0$. The layer of liquid metal is assumed to have an initial temperature distribution $f(x)$. It is essential for Miranker's formalism that A be positive and that $df(A)/dx$ be negative. Physically, this means that the front separating the liquid and solid metal must be moving before the mathematical model applies. The purpose of this paper is to present a constructive existence and uniqueness theorem which is not subject to this restriction.

The problem is to determine two functions $u(x, t)$ and $R(t)$ satisfying the following:

$$\begin{aligned} &u_{xx} = u_t\,, \qquad 0 < x < R(t), \qquad 0 < t, \\ &u_x(0, t) = -g(t), \qquad 0 < t, \\ &u(R(t), t) = -dR(t)/dt, \qquad 0 < t, \\ &R(0) = A, \\ &u(x, 0) = f(x), \; 0 < x < A, \end{aligned} \tag{1.1}$$

where g is a positive continuous function, and f is a continuous function such that, for some constant b,

$$0 \leq f(x) \leq b(A - x), \qquad 0 \leq x \leq A. \tag{1.2}$$

Miranker required that f be continuously differentiable, non-negative, and that

$$df(0)/dx = -g(0), \qquad df(A)/dx < 0, \tag{1.3}$$

the latter condition being essential for his proof. Although it was not stated explicitly, it follows from his conclusion that u_x is continuous on the boundary** that f must vanish at $x = A$. Hence Miranker's initial value function satisfies (1.2).

In 1951, G. W. Evans [3] published an existence theorem for this problem with $A = 0$ and g constant. His proof consisted of an iterative argument applied to a heat

*Received November 3, 1958. The research for this paper was done while the author was a Temporary Member of the Institute of Mathematical Sciences, New York University.

**See the proof of lemma 2 in [1].

balance equation. He proved the existence of a solution for t restricted to the interval $[0, 1/4]$. J. Douglas and T. M. Gallie [4], A. Datzeff [5], G. Sestini [6], A. Friedman [7], and the present author [8], have proved existence theorems for similar problems.

2. The existence theorem: There exists a unique solution to the free boundary problem.

Proof: Following Evans, we derive a heat balance equation by evaluating

$$\int_0^t \int_0^{R(t')} (u_{xx} - u_t)\, dx\, dt' = 0, \tag{2.1}$$

using the boundary conditions (1.1). We obtain

$$R(t) = A + \int_0^t g(s)\, ds - \int_0^{R(t)} u(x, t)\, dx + \int_0^A f(x)\, dx. \tag{2.2}$$

We use this equation to define a transformation $S = F(R)$ by taking R to be a given differentiable monotonic function such that $R(0) = A$, and taking u to be the solution of the *reduced problem*:

$$\begin{aligned} &u_{xx} = u_t\,, && 0 < x < R(t), \quad 0 < t, \\ &u_x(0, t) = -g(t), && 0 < t, \\ &u(R(t), t) = 0, && 0 < t, \\ &u(x, 0) = f(x), && 0 < x < A. \end{aligned} \tag{2.3}$$

If we can find a differentiable monotonic function which is left invariant by the transformation F, then it, together with the corresponding solution to the reduced problem, satisfies (1.1). In this paper, we show that the boundary function we seek is the limit function of a sequence of iterates, $R_0 = A$, $R_{n+1} = F(R_n)$.

The sequence of iterates is well defined, for if R is differentiable, and if $S = F(R)$, then S is monotonic and differentiable. In fact,

$$0 \leq dS(t)/dt = -u_x(R(t), t). \tag{2.4}$$

The equality follows from (2.2) and (2.1). The inequality from the following argument: if u_x were positive on $x = R(t)$, u would be negative nearby. But then, by the maximum principle, u would attain its negative minimum on $x = 0$. This cannot happen since u_x is negative there.

Our goal is to prove the existence of a solution for an arbitrary time interval. The first iterative process which we carry out will converge if the time interval is small. Then, taking as the initial function the solution to the reduced problem corresponding to the limit boundary function, we carry out another iteration with a small time step, etc. We show that a finite number of such processes will give the solution over an arbitrary time interval $[0, T]$. Uniqueness of the solution follows from the contracting character of the transformation F. If we are content with existence alone, we can use a standard fixed point theorem of functional analysis and obtain a (non-constructive) proof which does not require subdividing the time interval.*

We will show that the first iterative process converges to a differentiable function

*This method was used in [8]. Uniqueness was established by a separate argument.

if $0 < t \leq t_1$. Subsequent iterative processes provide solutions to the integral equations

$$R(t) = A_p + \int_{t_p}^{t} g(s)\, ds - \int_0^{R(t)} u(x, t)\, dx + \int_0^{A_p} u_p(x)\, dx, \tag{2.5}$$

where $A_p(= R(t_p))$ and $u_p(x)$ are obtained from the previous process. The function u is the solution to the reduced problem

$$\begin{aligned} &u_{xx} = u_t\,, \qquad 0 < x < R(t), \qquad t_p < t < t_{p+1}\,, \\ &u_x(0, t) = -g(t), \qquad t_p < t < t_{p+1}\,, \\ &u(R(t), t) = 0, \qquad t_p < t < t_{p+1}\,, \\ &u(x, 0) = u_p(x), \qquad 0 < x < A_p\,. \end{aligned} \tag{2.6}$$

In order to prove convergence, we need the following estimates:

Lemma 1. *If u is the solution to the reduced problem, then there exists a number B, independent of $R(t)$, such that for all t in the interval* $[0, T]$,

$$\begin{gathered} A \leq R(t) \leq Bt, \\ 0 \leq u(x, t) \leq B(R(t) - x), \qquad 0 \leq x \leq R(t), \\ -B \leq u_x(R(t), t) \leq 0. \end{gathered} \tag{2.7}$$

Lemma 2. *If u and v are solutions to the reduced problem* (2.6) *with boundary functions R and S respectively, then there exists $q_0 > 0$, independent of the boundary curves and of the subdivision, such that $0 < t - t_p < q_0$ implies that*

$$\int_0^{A_p} |\, u(x, t) - v(x, t)\,|\, dx \leq (1/2)\,|\, R - S|\,{}^{*}_{t_{p+1}}\,, \qquad t_{p+1} = t_p + q_0\,. \tag{2.8}$$

Furthermore, for all $t > t_p$,

$$|\, u(x, t) - v(x, t)| \leq B\,|\, R - S\,|^{*}_{t},\ 0 \leq x \leq \min\,(R(t), S(t)). \tag{2.9}$$

The derivation of these estimates is in the appendix.

Let

$$\begin{aligned} j_n(t) &= \min\,(R_n(t), R_{n-1}(t)), \\ k_n(t) &= \max\,(R_n(t), R_{n-1}(t)), \end{aligned} \tag{2.10}$$

then if $R_{n+1} = F(R_n)$ defines the pth iterative process, and if $q < q_0$,

$$\begin{aligned} &|\, R_{n+1} - R_n\,|_q \leq \int_0^{A_p} |\, \Delta u_n(x, t)\,|\, dx + \int_{A_p}^{j_n(t)} |\, \Delta u_n(x, t)\,|\, dx \\ &+ \int_{j_n(t)}^{k_n(t)} |\, \Delta u_n(x, t)\,|\, dx, \leq 1/2\,|\, R_n - R_{n-1}\,|_q + B\,|\, j_n(t) - A\,|\,|\, R_n - R_{n-1}\,|_q \\ &+ B/2\,|\, R_n - R_{n-1}\,|_q^2 \leq [1/2 + 2B^2 q]\,|\, R_n - R_{n-1}\,|_q\,, \end{aligned} \tag{2.11}$$

where Δu_n is the difference between the solutions to the reduced problems corresponding

---

*$|\, R - S\,|_t = \max\,|\, R(t') - S(t')\,|,\ t_p \leq t' \leq t$.

to R_n and R_{n-1} . We have adopted the convention that the solution to the reduced problem is identically zero outside the original domain of definition, i.e., $u(x, t) = 0$, if $x > R(t)$.

Clearly, if $q < \min(1/4B^2, q_0)$, the sequence converges to a monotonic function $R(t)$. Let $u(x, t; R)$ be the solution to the reduced problem corresponding to the limit function R. Then if $R' = F(R)$,

$$\begin{aligned} &| R' - R | = | F(R) - F(R_n) | + | R_{n+1} - R |, \\ &| R' - R |_t \leq [1/2 + 2B^2 t] | R - R_{n-1} |_t + | R_{n+1} - R |_t , \qquad 0 < t < q_1 . \end{aligned} \tag{2.12}$$

Since the right side can be made arbitrarily small, we conclude that R is invariant under F. We repeat this argument for each subinterval.

The functions $R(t)$ and $u(x, t; R)$ are the solution to the free boundary problem if R is differentiable. To prove that R is differentiable, we write

$$\begin{aligned} [R(t + k) - R(t)] + \int_{R(t)}^{R(t+k)} u(x, t)\, dx = \int_{R(t)}^{R(t+k)} g(s)\, ds \\ - \int_0^{R(t)} [u(x, t + k) - u(x, t)]\, dx. \end{aligned} \tag{2.13}$$

Using the law of the mean and the fact that u is the solution to the reduced problem, we get

$$(1/k)[R(t + k) - R(t)] = -u_x(R(t), t) + 0(k). \tag{2.14}$$

This concludes the proof of the theorem.

APPENDIX

Proof of Lemma 1. In proving inequality (2.4), we found that u_x must be non-positive on $x = R(t)$ and that u must be non-negative in the interior of the domain. To obtain the lower bound on u_x , we pick a constant B so that

$$\begin{aligned} &g(t) < B, \qquad 0 < t < T, \\ &0 \leq f(x) \leq B(A - x), \qquad 0 \leq x \leq A. \end{aligned} \tag{a1}$$

We extend f as an even function and take v to be the solution of the heat equation taking on the boundary values

$$\begin{aligned} &v(x, 0) = f(x) + B | x |, \qquad -A < x < A, \\ &v(\pm R(t), t) = BR(t), \qquad 0 < t < T. \end{aligned} \tag{a2}$$

$R(t)$ is monotonic, so by the maximum principle,

$$\begin{aligned} &0 \leq v(x, t) \leq BR(t), \\ &0 \leq v_x(R(t), t). \end{aligned} \tag{a3}$$

Since v is an even function, $v_x = 0$ on $x = 0$. Hence, if we let

$$w(x, t) = v(x, t) - Bx - u(x, t), \tag{a4}$$

then

$$w_x = v_x - B - u_x < 0 \quad \text{on} \quad x = 0, \qquad 0 < t < T. \tag{a5}$$

By construction, $w = 0$ on $x = R(t)$. It follows from the maximum principle that

$$\begin{gathered} 0 \leq w(x, t), \qquad 0 \leq x \leq R(t), \qquad 0 < t < T, \\ w_x(R(t), t) \leq 0. \end{gathered} \tag{a6}$$

We conclude that

$$\begin{gathered} -B \leq v_x(R(t), t) - B \leq u_x(R(t), t), \qquad 0 < t < T, \\ u(x, t) \leq v(x, t) - Bx \leq B(R(t) - x), \qquad 0 \leq x \leq R(t), \qquad 0 < t < T. \end{gathered} \tag{a7}$$

Proof of Lemma 2. It follows from (2.7) that (2.9) is valid on the boundary, $x = R(t)$. By the maximum principle, it is valid in the interior of the domain.

If $R \not\equiv S$, let $t' = \sup \{t \mid R(t) = S(t)\}$. Then for any $r > 0$,

$$\begin{gathered} | u(x, t) - v(x, t) | \leq w(x, t) \, | R - S |_{t'+r}, \; t' < t < t' + r, \\ 0 \leq x \leq C = R(t') = S(t'), \end{gathered} \tag{a8}$$

where w is the solution to

$$\begin{aligned} &w_t = w_{xx}, && 0 < x < C, && t' < t, \\ &w(x, t') = 0, && 0 < x < C, \\ &w(C, t) = B, && t' < t, \\ &w_x(0, t) = 0, && t' < t. \end{aligned} \tag{a9}$$

Note that

$$A \leq C \leq BT, \qquad 0 \leq t' < T. \tag{a10}$$

Clearly,

$$\int_0^C | u(x, t) - v(x, t) | \, dx \leq | R - S |_{t'+r} \int_0^C w(x, t) \, dx. \tag{a11}$$

Since

$$\int_0^C w(x, t) \, dx = 4C \sum_{n=0}^{\infty} (1/(2n + 1)^2(1 - \exp - \{(t - t')(2n + 1)^2/4C^2\}) \tag{a12}$$

approaches zero uniformly in C and t' as t approaches t', we can restrict r so that the integral (a12) is less than $1/2$.

If $A = 0$, the estimate (2.8) is not needed for the first iterative process. The lower bound on C will then be $R(t_1)$.

Bibliography

1. W. L. Miranker, *A free boundary value problem for the heat equation*, Quart. Appl. Math. **16**, 121–130 (1958)
2. I. I. Kolodner, *Free boundary problem for the heat equation with applications to problems of change of phase*, Communs. Pure Appl. Math. **9**, 1–31 (1956)

3. G. W. Evans, *A note on the existence of a solution to a problem of Stefan*, Quart. Appl. Math. 9, 185–193 (1951)
4. J. Douglas and T. M. Gallie, *On the numerical integration of a parabolic differential equation subject to a moving boundary condition*, Duke Math. J. 22, 557–572 (1955)
5. A. Datzeff, *Sur la probleme lineaire de Stefan*, Annuaire univ. Sofia, Livre I, 46, 271–325 (1950)
6. G. Sestini, *Esistenza di una soluzione in problemi analogli a quella di Stefan*, Riv. Matematica Univ. Parma 3, 171–180 (1929)
7. A. Friedman, *Free boundary problems for parabolic equations*, University of California (Berkeley) Tech. Rept. No. 28 (Oct. 1958)
8. W. T. Kyner, *An existence and uniqueness theorem for a nonlinear Stefan problem*, J. Math. Mech. (in press)

—NOTES—

ON A SPECIAL BOLZA VARIATIONAL PROBLEM, AND THE MINIMIZATION OF SUPERAERODYNAMIC HYPERSONIC NOSE DRAG*

By H. S. TAN (*Therm-Electric Meters Co., Ithaca, N. Y.*)

Instead of a definite integral, let the following expression be given:

$$D = G(y_0 , y_l) + \int_0^l F(y, y', x)\, dx, \tag{1}$$

where G is a known function of y_0 and y_l , y_0 and y_l may or may not be specified. It is desired to find the optimum function $y(x)$ that minimizes expression (1). This is a special case of the Bolza problem [1].

The variation of (1) is easily obtained as follows:

$$\begin{aligned}\delta D &= D(y^*) - D(y) \\ &= G_{y_0}\, \delta y_0 + G_{y_l}\, \delta y_l + [F_{y'}\, \delta y]_0^l + \int_0^l [F_y - (F_{y'})']\, \delta y\, dx \qquad (2) \\ &= [G_{y_0} - (F_{y'})_0]\, \delta y_0 + [G_{y_l} + (F_{y'})_l]\, \delta y_l + \int_0^l [F_y - (F_{y'})']\, \delta y\, dx.\end{aligned}$$

To insure vanishing of δD, it is clear that:

(i) throughout the interval $0 < x < l$, Euler's variational equation must be satisfied:

$$\varphi = (F_{y'})' - F_y = 0 \tag{3}$$

(ii) at both ends of the interval, the following end conditions must be met:

$$\begin{aligned}\delta y_0 = 0, \quad \text{or} \quad \psi_0 = G_{y_0} - (F_{y'})_0 = 0, \\ \delta y_l = 0, \quad \text{or} \quad \psi_l = G_{y_l} + (F_{y'})_l = 0.\end{aligned} \tag{4}$$

The condition of minimization is then furnished by

(iii) the Legendre second variational inequality:

$$F_{y'y'} > 0, \qquad 0 < x < l. \tag{5}$$

Our generalization evidently lies in the inclusion of G term outside the integral, and the corresponding broadened end conditions (4). Indeed, if G is constant, or disappears from (1), then the end conditions simply reduce to the following conventional form for calculus of variations [2, 3, 4], i.e.

$$\begin{aligned}\delta y_0 = 0, \quad \text{or} \quad (F_{y'})_0 = 0, \\ \delta y_l = 0, \quad \text{or} \quad (F_{y'})_l = 0.\end{aligned} \tag{6}$$

---

* Received June 9, 1958.

In case

$$G(y_0 , y_l) = \int_{0-}^{0+} F(y, y', x)\, dx + \int_{l-}^{l+} F(y, y', x)\, dx,$$

(1) reduces to following single integral:

$$D = \int_{0-}^{l+} F(y, y', x)\, dx \tag{7}$$

which is, on the one hand, an immediate extension of the ordinary calculus of variations to include Stieltjes integral at the ends; and on the other hand, a special case of our present formulation.

The present generalization of the variational problem has, in fact, arisen as a result of the search for optimum nose curve that minimizes the superaerodynamic hypersonic drag of an axially symmetric body. The problem is usually so formulated that the nose length l and base radius y_l are both given, while the tip radius y_0 may or may not be specified, depending on the situation. By using meridianial coordinates, x axial and y radial, the nose drag has been shown to be given by following integral [5]:

$$\begin{aligned} D &= 2\pi\rho V^2 \int_{0-}^{l} \{1 + cy'(1 + y'^2)^{-1/2}\} yy'\, dx \\ &= \pi\rho V^2 \left\{ y_l^2 + c\left[y_0^2 + 2\int_0^l yy'(1 + y'^2)^{-1/2}\, dx \right] \right\}, \end{aligned} \tag{8}$$

where c is a constant determined by the ratios of solid surface and gas temperature, and of body and molecular speed. It is easy to see that this drag formula is indeed of type (1), with

$$G = y_0^2 + (y_l^2/c), \qquad F = 2yy'(1 + y'^2)^{-1/2} \tag{9}$$

and end conditions:

$$\begin{aligned} \delta y_l &= 0 \\ \delta y_0 = 0, \quad \text{or} \quad \psi_0 &= G_{y_0} - (F_{y'})_0 = 0. \end{aligned} \tag{10}$$

As already pointed out, an optimum solution must be that of Euler's variational equation, which, by putting (9) into (3), takes the following form:

$$y'' = -y'^2(1 + y'^2)/y(2 - y'^2). \tag{11}$$

Its solution, in parametric form, has been obtained, by simple quadrature and use of relations $y'' = y'dy'/dy$ and $y'' = dy'/dx$, as follows [6]:

$$\begin{aligned} y &= c_1(1 + y'^2)^{3/2}/y'^2, \\ x &= \frac{c_1}{3}\left\{ \frac{2(1 + y'^2)^{1/2}}{y'^3} - \frac{(1 + y'^2)^{1/2}}{y'} + 3 \ln [(1 + y'^2)^{1/2} + y'] \right\} + c_2 . \end{aligned} \tag{12}$$

On putting (9) into (5), the second variational inequality then requires:

$$F_{y'y'} = 2 - y'^2 > 0, \quad \text{i.e.} \quad y' < (2)^{1/2} \tag{13}$$

Differential equation (11) shows there are two branch curves merging at cuspidal

point $y' = (2)^{1/2}$; solution (12) shows that the axis $y = 0$ can not be reached, and y grows indefinitely both as y' approaches zero and infinity. The correct branch of curve is thus specified by requirement (13) which, together with (12), imposes the following limitations on the admissible values of y':

$$0 < y' < (2)^{1/2}, \qquad y'' < 0. \tag{14}$$

Now the tip condition $\psi_0 \delta y_0 = 0$ becomes:

(i) For specified y_0: $\delta y_0 = 0$. In this case, in view of the limitations on the admissible values of y', it is evident that with prescribed nose length l and base radius y_l, there is a lower bound for admissible tip radius y_0, below which no solution exists. Within its admissible range, it is easy to see that specification of x_0, y_0, x_l, y_l, uniquely determines the four unknowns C_1, C_2, y_0' and y_l', by the four independent equations from parametric solution (12).

(ii) For unspecified y_0:

$$\psi_0 = G_{y_0} - (F_{y'})_0 = y_0[1 - y_0'(2 + y_0'^2)(1 + y_0'^2)^{-3/2}] = 0. \tag{15}$$

Although it is easy to see that both $y_0 = 0$ and $y_0' = \infty$ are solutions of (15), neither of them can be fulfilled by our solution (12), with finite l and non-vanishing y_l. However, a plot of ψ_0 against y_0' (Fig. 1) shows that the curve crosses axis $\psi_0 = 0$ at $y_0' = 0.7862 < (2)^{1/2}$ Thus, with unspecified y_0, y_0' can be determined through condition (15). Actual construction of the optimum curve then amounts to specifying x_0, x_l, y_0' (through $\psi_0 = 0$), and y_l, from which four unknowns y_0, y_l', C_1, C_2, are determined by four independent equations from parametric solution (12).

It is interesting to note that in this case y_0 corresponding to $y_0' = 0.7862$ is usually greater than the least admissible value of y_0 which corresponds to $y_0' = (2)^{1/2}$. This implies that with a given finite nose length l, reducing the tip radius beyond y_0 corresponding to $y_0' = 0.7862$ actually has an adverse effect. To see this point, it is best to

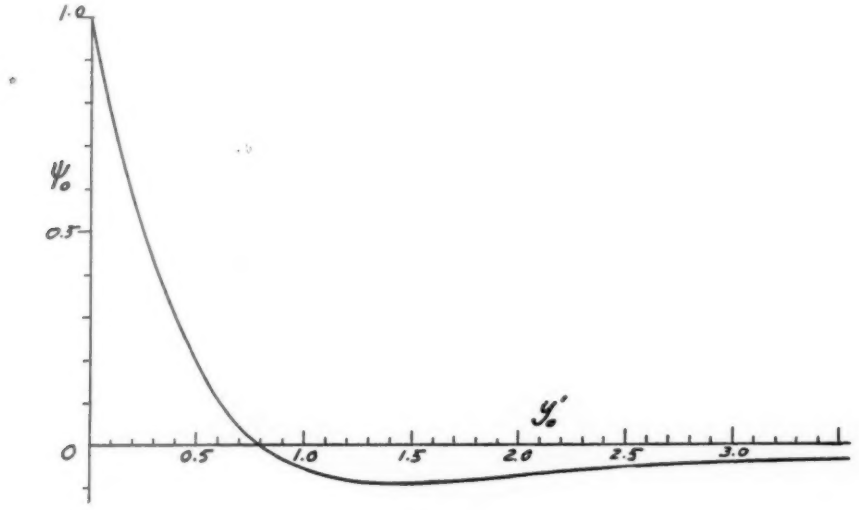


FIG. 1.

refer to Fig. 1, which shows clearly that both for $y_0(y_0')$ greater and smaller than y_0 (.7862), we will have $\psi_0 \delta y_0 > 0$, i.e., $\delta D > 0$.

References

1. G. A. Bliss, *Lectures on the calculus of variations*, University of Chicago Press, 1947
2. O. Bolza, *Lectures on the calculus of variations*, University of Chicago Press, 1904
3. G. A. Bliss, *Calculus of variations*, Open Court Co., 1925
4. H. and B. S. Jeffreys, *Mathematical physics*, Cambridge University Press, 1950
5. H. S. Tan, *On superaerodynamic drag, and its minimization*, to be published
6. H. S. Tan, *On optimum nose curves for superaerodynamic missiles*, J. Aeronaut. Sci. 25, 263 (1958)

NOTE ON THE SOLUTION OF THE NEUTRON DIFFUSION PROBLEM BY AN IMPLICIT NUMERICAL METHOD*

By GEORGE A. BAKER, JR. (*Los Alamos Scientific Laboratory, Los Alamos, New Mexico*)

It is of interest that the implicit, numerical method of Baker and Oliphant [1, 2] for the solution of time-dependent heat-flow problems in rectangular regions may be usefully extended to obtain the solution of the time-independent neutron-diffusion equation

$$\nabla D(r) \nabla \varphi - \chi(r)\varphi + S(r) = 0 \tag{1}$$

for the neutron density φ. In the original work of [1], to obtain an accurate solution, even asymptotically, it was necessary to take Δt small. This choice was necessary because of the occurrence of a term in a higher order space derivative multiplied by Δt. We will herein make a slight modification to remove this defect. Let us rewrite (1) as (for the two-dimensional case)

$$\begin{aligned} \beta\varphi + \nabla^2\varphi + \frac{1}{\beta}\frac{\partial^4\varphi}{\partial x^2\,\partial y^2} = \alpha^{-2}\frac{\partial\varphi}{\partial t} + \beta\varphi + \frac{1}{\beta}\frac{\partial^4\varphi^*}{\partial x^2\,\partial y^2} \\ + \left(\frac{\chi(r)}{D(r)} - \mu\right)\varphi^* + \mu\varphi - \frac{S(r)}{D(r)} + \frac{\nabla D\cdot\nabla\varphi^*}{D(r)}, \end{aligned} \tag{2}$$

where

$$\varphi = \varphi^*, \qquad \frac{\partial\varphi}{\partial t} = 0, \tag{3}$$

and

$$\mu = \operatorname*{Min}_r\,[\chi(r)/D(r)]. \tag{4}$$

To obtain the solution of (1) via (2) and (3) we must advance the time until the asymptotic solution of (2) is obtained. In [1], we described for the special case, D, a constant, how to guess φ^*, and then use (2) to calculate φ_{calc}. by solving the left-hand side and then how to compute a new guess, φ^{**}, by means of

*Received September 30, 1958. Work performed under the auspices of the United States Atomic Energy Commission.

$$\varphi^{**} = \frac{(F + \omega - 1)\varphi^* + \varphi_{\text{calc.}}}{F + \omega}, \tag{5}$$

where

$$\lambda = \frac{-3\alpha^{-2}}{2\,\Delta t}, \beta = \lambda - \mu, \quad \omega = \lambda/\beta,$$
$$F = -\chi(r)/[D(r)\beta], \qquad W = -S(r)/[D(r)\beta]. \tag{6}$$

In [1] this process was continued until φ^{**} converged within a desired accuracy to φ. For the steady state, if we obtain φ^* by linear extrapolation from the two previous times, then the first φ^{**} obtained is a sufficiently good approximation for φ and one can proceed directly to the next time step.

If we assume that $D(r)$ half-way between mesh points is obtained by averaging the reciprocals (for conservation of flux at discontinuities), then the proper difference equation representation for $\beta(\Delta x)^2(\Delta y)^2$ times the right-hand side of (2) would be

$$\begin{aligned}
&\beta(\Delta x)^2\beta(\Delta y)^2\{W(i,j) + \omega[\tfrac{4}{3}\,{}^1\varphi(i,j) - \tfrac{1}{3}\,{}^2\varphi(i,j)] - [F(i,j) + \omega - 1]\varphi^*(i,j)\} \\
&\quad - \beta(\Delta y)^2\left\{\left[\frac{D(i+1,j) - D(i,j)}{D(i+1,j) + D(i,j)}\right][\varphi^*(i+1,j) - \varphi^*(i,j)]\right. \\
&\quad \left. + \left[\frac{D(i,j) - D(i-1,j)}{D(i,j) + D(i-1,j)}\right][\varphi^*(i,j) - \varphi^*(i-1,j)]\right\} \\
&\quad - \beta(\Delta x)^2\left\{\left[\frac{D(i,j+1) - D(i,j)}{D(i,j+1) + D(i,j)}\right][\varphi^*(i,j+1) - \varphi^*(i,j)]\right\} \\
&\quad + \left[\frac{D(i,j) - D(i,j-1)}{D(i,j) + D(i,j-1)}\right][\varphi^*(i,j) - \varphi^*(i,j-1)] \\
&\quad + 4\varphi^*(i,j) - 2[\varphi^*(i+1,j) + \varphi^*(i,j+1) \\
&\quad + \varphi^*(i-1,j) + \varphi^*(i,j-1)] + \varphi^*(i+1,j+1) \\
&\quad + \varphi^*(i+1,j-1) + \varphi^*(i-1,j+1) + \varphi^*(i-1,j-1),
\end{aligned} \tag{7}$$

where ${}^1\varphi$ is φ at $t - \Delta t$ and ${}^2\varphi$ is φ at $t - 2\Delta t$.

In order to obtain the asymptotic solution as quickly as possible, we would like to choose Δt as large as possible. Hence we wish to pick $-\beta$ as small as possible. However, as was pointed out in [1], the method loses about $[-\log_2(0.2\,\beta^2(\Delta x)^2(\Delta y)^2)]$ binary bits due to the cancellation of nearly equal numbers. Therefore, if we carry eight decimal places and wish to retain four figure accuracy, we must restrict

$$\beta^2(\Delta x)^2(\Delta y)^2 \gtrsim 5 \times 10^{-4}. \tag{8}$$

As a general rule, our experience has shown that about 5 to 10 iterations are required and the choice of a smaller β than that needed to assure convergence in this number of iterations does not speed up the convergence. We find

$$\beta(\Delta x)^2 \quad \text{and} \quad \beta(\Delta y)^2 \sim -(\mu + 50/N), \tag{9}$$

where N is the number of mesh points, to be a good rough guide to the size of β to be used, subject of course to (8).

When $D(r)$ is not constant, β is subject to another restriction besides (8). If we represent φ^*, $\varphi_{\text{calc.}}$, and φ^{**} as φ plus an error term, it is easy to show using the method of Sec. 3 of [1] that if ϵ^* is the error in φ^*, then the error in φ^{**}, ϵ^{**}, is to first order approximately

$$\epsilon^{**} = -\frac{\nabla D \cdot \nabla \epsilon^*}{\beta D(F + \omega)}. \tag{10}$$

From (7) it is evident that the x, and y components of $(\nabla D/D)$ are bounded in magnitude by $(\Delta x)^{-1}$ and $(\Delta y)^{-1}$ respectively. The term $F + \omega$ has a minimum value of unity. At a discontinuity (Δx or Δy) $\nabla \epsilon^*$ may well be of the order of ϵ^*, hence we see $-\beta(\Delta x)^2$ and $-\beta(\Delta y)^2$ must be at least unity. We find in practice that a satisfactory choice of β may be obtained from the rule that $-\beta(\Delta x)^2$ and $-\beta(\Delta y)^2$ are greater than 2. If the discontinuities in D are slight then we can reduce the value of β correspondingly. It seems desirable to provide a convergence factor of about $\frac{1}{2}$ between ϵ^* and ϵ^{**} in (10).

In the problems ($N \lesssim 2000$) we have run using this method, most took about 7 to 15 iterations to reduce the error by a factor of 100. Large problems with sharp discontinuities in D can take considerably longer, due to the restriction imposed by (10) on β.

References

1. G. A. Baker, Jr. and T. A. Oliphant, *An implicit, numerical method for solving the two dimensional heat equation*, to appear in Quart. Appl. Math.
2. G. A. Baker, Jr., *An implicit, numerical method for solving the n-dimensional heat equation*, to appear in Quart. Appl. Math.

FUNCTIONAL EQUATIONS AND MAXIMUM RANGE*

By RICHARD BELLMAN (*The RAND Corporation*)

1. Introduction. The current interest in rockets and space travel has aroused a corresponding interest in the determination of maximum range, minimum time, and so on, for various types of trajectories.

A variety of questions of this type have been treated by means of the theory of dynamic programming, see [1, 2, 4]. Here we wish to show how to use functional equations to determine the range, the maximum elevation, and similar quantities, as functions of initial position and velocities.

2. Vertical motion—I. Consider an object, subject only to the force of gravity and the resistance of the air, which is propelled straight up. In order to illustrate the technique we shall employ, let us treat the problem of determining the maximum altitude.

Let the defining equation be

$$u'' = -g - h(u'), \tag{1}$$

with the initial conditions $u(0) = 0$, $u'(0) = v$. Here $v > 0$, and $h(u') \geq 0$ for all u'.

Since the maximum altitude is a function of v, let us introduce the function

$$f(v) = \text{the maximum altitude attained starting with initial velocity } v. \tag{2}$$

*Received October 20, 1958.

From the definition of the function it follows that

$$f(v) = v\,\Delta + f(v - [g + h(v)]\,\Delta) + o(\Delta), \tag{3}$$

for Δ an infinitesimal. Verbally, this states that the maximum altitude is the altitude gained over an initial time Δ, plus the maximum altitude attained starting with a velocity $v - [g + h(v)]\Delta$, the velocity of the object at the end of time Δ, to within $o(\Delta)$.

Expanding both sides and letting $\Delta \to 0$, we see that

$$f'(v) = \frac{v}{g + h(v)}. \tag{4}$$

Since $f(0) = 0$, this yields

$$f(v) = \int_0^v \frac{v_1\,dv_1}{g + h(v_1)}. \tag{5}$$

In the particular case where $h(v) = 0$, we obtain the standard result $v^2/2g$.

3. Vertical motion—II. Consider the more general case where motion is through an inhomogeneous medium. Let the defining equation be

$$u'' = h(u, u'), \qquad u(0) = c_1\,, \qquad u'(0) = c_2\,. \tag{1}$$

Assume that $h(u, u') \leq 0$ for all u and u', so that $c_2 = 0$ implies no motion.

The maximum altitude is now a function of both c_1 and c_2. Introduce

$$f(c_1\,, c_2) = \text{the maximum altitude attained starting with the initial position } c_1 \text{ and initial velocity } c_2\,. \tag{2}$$

Then, as above,

$$f(c_1\,, c_2) = c_2\,\Delta + f[c_1 + c_2\,\Delta, c_2 + h(c_1\,, c_2)\,\Delta] + o(\Delta), \tag{3}$$

which yields in the limit the partial differential equation

$$c_2 + c_2\frac{\partial f}{\partial c_1} + h(c_1\,, c_2)\frac{\partial f}{\partial c_2} = 0. \tag{4}$$

By virtue of our assumptions, $f(c_1\,, 0) \equiv 0$, for $c_1 \geq 0$.

4. Computational aspects. One can, of course, use the method of characteristics, or standard difference methods, to solve (3.4). Let us present another method which reduces the solution to the tabulation of a sequence of functions of one variable.

In place of (3.4), let us use the discrete approximation of (3.3),

$$f(c_1\,, c_2) = c_2\,\Delta + f[c_1 + c_2\,\Delta, c_2 + h(c_1\,, c_2)\,\Delta]. \tag{1}$$

Since c_2 is monotone decreasing, it can be used to play the role of time. Let us write $c_2 = N\delta$, where δ is a positive quantity, and $f(c_1\,, c_2) \equiv f_N(c_1)$. We consider then only values of c_2 which are multiples of δ. To overcome the fact that $c_2 + h(c_1\,, c_2)\Delta$ in general will not be a multiple of δ, we can either replace it by $[(c_2 + h(c_1\,, c_2)\Delta)/\delta]$, or use interpolation. Although use of an interpolation formula slows up the computation, it greatly improves the accuracy. For an application of the foregoing techniques to a more complicated partial differential equation, see [3].

5. Maximum altitude. Consider now the case where motion takes place in a plane. Let the equations be

$$x'' = g(x', y'), \qquad x(0) = 0, \qquad x'(0) = c_1 , \tag{1}$$
$$y'' = h(x', y'), \qquad y(0) = 0, \qquad y'(0) = c_2 .$$

Introducing, as before, the function $f(c_1 , c_2)$ equal to the maximum altitude, we see that

$$f(c_1 , c_2) = (c_1^2 + c_2^2)^{1/2} \Delta + f[c_1 + g(c_1 , c_2) \Delta, c_2 + h(c_1 , c_2) \Delta] + o(\Delta). \tag{2}$$

Hence,

$$(c_1^2 + c_2^2)^{1/2} + g(c_1 , c_2) \frac{\partial f}{\partial c_1} + h(c_1 , c_2) \frac{\partial f}{\partial c_2} = 0. \tag{3}$$

Once again, let us assume that $c_2 = 0$ implies no vertical motion. Then $f(c_1 , 0) = 0$ for $c_1 \geq 0$. It follows that we can again compute the solution by means of a sequence of functions of one variable.

6. Maximum range. To tackle the problem of maximum range directly requires the introduction of another state variable, the initial altitude. It can also be broken up into two problems, corresponding to the ascent to maximum altitude, and the descent.

References

1. R. Bellman, *Dynamic programming*, Princeton University Press, Princeton, N. J., 1957
2. R. Bellman and S. Dreyfus, *An application of dynamic programming to the determination of optimal satellite trajectories*, to appear in J. Brit. Interplanetary Soc.
3. R. Bellman, I. Cherry, and G. M. Wing, *A note on the numerical integration of a class of nonlinear hyperbolic equations*, Quart. Appl. Math. **16**, 181-183 (1958)
4. T. Cartaino and S. Dreyfus, *Application of dynamic programming to the airplane minimum time-to-climb problem*, Aeronaut. Eng. Rev. **16**, 74-77 (1957)

ON THE DETERMINATION OF CERTAIN THERMODYNAMIC AND PHYSICAL QUANTITIES*

By A. GLEYZAL (*U. S. Naval Ordnance Laboratory, White Oak, Silver Spring, Maryland*)

We consider any physical phenomenon where a quantity z is a continuous differentiable function of two independent quantities x and y. Thus:

$$z = z(x, y).$$

Hence

$$dz = F\,dx + G\,dy,$$

where

$$F = F(x, y) = \frac{\partial z}{\partial x},$$

$$G = G(x, y) = \frac{\partial z}{\partial y}.$$

*Received October 27, 1958.

Suppose furthermore that the quantities G and y are readily measured directly but F and x cannot be measured directly except that the family of curves $F = \text{const.}$ may be determined and at least two curves $x = \text{const.}$ may be found. Then the physical quantities F and x themselves may be determined as functions of G and y. We need merely assign a label F to each curve of the family of curves $F = \text{const.}$ in such a manner that the areas enclosed by the two curves $x = \text{const.}$ and two curves $F = \text{const.}$ are proportional to the increment ΔF of F. Then additional curves $x = \text{const.}$ may be constructed and x evaluated so that the area enclosed by the two curves $x = x_1$, $x = x_1 + \Delta x$, and two curves $F = F_1$ and $F = F_1 + \Delta F$ is equal to $\Delta F \Delta x$.

Proof. Given any exact differential

$$dz = F\,dx + G\,dy, \tag{1}$$

where F and G are functions of x and y, then, if the curves $F = \text{const.}$ and $x = \text{const.}$ are plotted in the G, y plane as shown in Fig. 1,
the area ΔA enclosed by the curves:

$$F = F_1 , \qquad F = F_1 + \Delta F,$$

$$x = x_1 , \qquad x = x_1 + \Delta x,$$

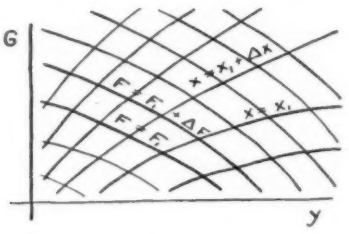


Fig. 1.

is equal to $\Delta F \Delta x$. For, integrating

$$\int_c dz = \int_c F\,dx + \int_c G\,dy = 0,$$

where C is the cycle which proceeds along curves $x = \text{const.}$ or $F = \text{const.}$ from x_1 , F_1 to x_1 , $F_1 + \Delta F$, to $x_1 + \Delta x$, $F_1 + \Delta F$, to $x_1 + \Delta x$, F_1 to x_1 , F_1 , we find:

$$0 = \Delta F\,\Delta x + \int_c G\,dy.$$

Therefore:

$$\int_c G\,dy = \Delta A = -\Delta F\,\Delta x.$$

These statements may be generalized to three or more variables and are themselves generalizations of, and render obvious, certain relationships among thermodynamic variables. Here, in the usual notation, since

$$dE = T\,dS - p\,dV, \tag{2}$$

we have

$$-\int_C p\,dV = \Delta A = -\Delta S\,\Delta T, \tag{3}$$

where C is a Carnot cycle which proceeds first along an adiabatic, then along an isothermal, then an adiabatic, and then along an isothermal to the starting point. We conclude that the family of isothermals $T = T_1 + n\Delta T$ and the family of adiabatics $S = S_1 + n\Delta S$, $n = 0, \pm 1, \pm 2, \pm 3, \cdots$ drawn in the p, V plane map out equal areas ΔA in this plane. Thus, if for any gas the isothermals $p = p_i(V)$ and the adiabatics $p = p_a(V)$ are determined by experiment, the "labels" S and T for the curves may then be determined for it is merely necessary to label as 0 the isothermal along which water freezes and as 100 the isothermal along which water boils at atmospheric pressure. The labels of the curves T = const. are then determined by Eq. (3) uniquely. Any pair of curves may be labeled $S = 1$ and $S = 2$ along any isothermal. The unit of entropy is related to the unit of mechanical energy by resorting again to Eq. (3). The labels for the intervening curves S = const. are also uniquely determined by Eq. (3). Entropy and absolute temperature S and T as thus determined for one gas must be consistent with S and T determined for any other gas due to the principle of conservation of energy. The extent to which the areas ΔA in the p, V plane are equal indicates the extent to which the postulated equation (2) is valid.

Correction to my paper

ON TRANSFER FUNCTIONS AND TRANSIENTS

Quarterly of Applied Mathematics, XVI, 273-294 (1958)

By A. H. ZEMANIAN (*New York University*)

Expression (38) should be replaced by its $(m - n)$th positive root and the second line below this expression should read, "horizontal line whose ordinate is the $(m - n)$th positive root of $(m - n)$!."

The second conclusion of Theorem 8 should be deleted and its proof adjusted such that zeros on the imaginary axis are counted with the zeros in the left half plane (i.e. the symbol q should be discarded and the expression $n - p - q$ should be replaced by $n - p$).

BOOK REVIEWS

(*Continued from p. 298*)

Some aspects of the mathematical theory of control processes. By R. E. Bellman, I. Glicksberg, O. A. Gross. The RAND Corp., Santa Monica, California, 1958. xix + 244 pp.

The purpose of the book is to provide "a taste of the mathematical theory of control processes", and this has been accomplished by a blending of techniques and considerable mixing of a variety of topics. The authors selected the ingredients from their own contributions while at the same time giving references to more complete discussions and to the work of others. The book is far superior to and much more interesting than a survey. The authors have purposely not gone deeply into the mathematical theory nor have they emphasized applications. The reader interested in either of these aspects can consult the references, and the authors so advise him.

The most interesting feature of the book is the emphasis placed on the formulation of control problems and the techniques available for their solution and the illustration that by another formulation of the same problem new techniques become applicable. The techniques are those of differential equations, variational calculus, linear (Hilbert) spaces, dynamic programming and game theory.

A fairly general description of the control problem—sufficiently general to encompass problems in control of interest to analysts and operations researchers, economists, management consultants, and engineers—is given. The state of a physical system is described at time t by an n-dimensional vector

$$x(t) = (x_1(t), x_2(t), \cdots, x_n(t)).$$

The rate of change of the state of the system is described by an ordinary differential equation

$$\frac{dx}{dt} = G(x(t), f(t)), \qquad x(0) = c,$$

where the freedom in the choice of the control function f represents our ability to control the system, and we wish to improve the performance of the system by a propitious choice of f. The performance of the system under control f is measured by a functional $J(f)$, and an optimal choice of the control function f is that which minimizes (or maximizes) $J(f)$.

Although it is impossible here to describe all of the various control problems discussed in this book, we can describe the three principal problems. In each of these the differential equation is assumed to be linear with constant coefficients—

$$\frac{dx}{dt} = Ax + f, \qquad x(0) = c.$$

A is a constant matrix—and the control enters as a forcing term. The state x of the system (the output) is then linearly related to the control f (the input) by

$$x(t) = y(t) + \int_0^t K(t - s)f(s)\, ds,$$

where y is the solution of the homogeneous equation (the uncontrolled system). The first control problem is one where the functional J and the constraints on the allowable control functions are linear. For instance,

$$J(f) = \int_0^T (a, x)\, dt\,,$$

$$0 \leqq f_i(t) \leqq m_i\,, \qquad 0 \leqq t \leqq T, \qquad i = 1, \cdots, n,$$

and

$$\int_0^T (f, b)\, dt \leqq k\,,$$

where

$$(y, z) = \sum_{i=1}^{n} y_i z_i$$

is the inner product.

In the second control problem the functional is nonlinear and there are no constraints. An example of a nonlinear (quadratic) functional is

$$J(f) = \int_0^T (x - c, x - c)\, dt + a \int_0^T (f, f)\, dt.$$

This can be interpreted as a desire to maintain the system in the state c during the time interval $[0,T]$. The first term represents the cost of deviation from the state c, and the second term measures the cost of control. Optimal control minimizes the cost $J(f)$. The third control problem is one in which the functional J is nonlinear and there are linear constraints.

Part I of the book is a survey of fundamental results on linear functional equations (primarily, difference and differential equations). Part II is concerned with control problems of the first and second kind. A theoretical solution of a problem of the first kind is obtained using the Neyman-Pearson lemma. The techniques involved in computing the optimal control functions are illustrated. Problems of the second kind are solved by Hilbert-space techniques. Here again computation of a solution is discussed. Part III is a study of three problems of the third kind each attacked differently. First a variant of the Neyman-Pearson lemma is used, and next a problem with constraints is solved by a combination of classical techniques and *ad hoc* methods. The third problem is the "bang-bang" control problem and is solved by methods which utilize techniques of classical differential equations, linear spaces and dynamic programming. Their solution of the problem provides information on how to determine optimal switching in a bang-bang system, and this is illustrated for a two-dimensional system. In this example we see an aspect of the control problem that this book neglects. The practical problem—and this is certainly true of a servomechanism—is to determine the control function, not as a function of time, but as a function of the state of the system. In this case their methods give such a solution.

In Part IV there is an introduction to the theory of dynamic programming which gives a general formulation of both deterministic and stochastic multistage decision processes. The variational problem with constraints, which was previously treated by classical techniques, is now considered to be a multistage decision process. The discrete version of the continuous process provides a means of studying the structure of the solution and gives a numerical method of computation. Part V corresponds to Part IV with game theory in place of dynamic programming. The min-max technique of game theory is applied to some control problems with nonanalytic functionals J.

The authors provide, as they say, a taste of some aspects of control theory in which they have specialized and do so remarkably well. The book contains some misprints, some errors and a few false statements, none of which should cause the intelligent reader great concern. They have covered a wide variety of topics in a vast and exciting field of research with great skill.

J. P. LaSalle

Mathematical theory of compressible fluid flow. By Richard von Mises. Completed by Hilda Geiringer and G. S. S. Ludford. Academic Press, Inc., New York, 1958. xiii + 514 pp. $15.00.

Hilda Geiringer (Mrs. R. von Mises) and G. S. S. Ludford have done the scientific community a great service in completing for publication the last work in compressible flow of Professor von Mises. Unfortunately the present text is only the first part of what the author had originally intended to be a comprehensive work on compressible flow. However, this cannot be considered to detract from what has been published, but rather to leave uncovered by Professor von Mises' unique approach various topics of importance in compressible flow. The book is written principally from the applied mathematical point of view and is divided into five chapters, of which the first three were written by the author and the last two were completed according to his plan following his papers and lecture notes.

In the first chapter the momentum and energy equations are derived including the effect of viscous

stresses and heat conduction. Although the concept of a general specifying equation of state is introduced most of the special cases refer to a perfect gas. The chapter concludes with a discussion of the propagation of small disturbances in an inviscid fluid, along with a delineation of subsonic and supersonic motion. The second chapter presents general theorems of fluid motion and the method of characteristics, with principal emphasis on the mathematical aspects of problems in two independent variables.

The third chapter considers one-dimensional flow, with the first section treating the steady case including viscosity and heat conduction, while succeeding sections deal with the nonsteady flow of an ideal fluid. The chapter concludes on a discussion of shock phenomena with the treatment restricted to a perfect gas of constant specific heat ratio. In the fourth chapter the author deals with plane steady potential flow and discusses the hodograph method, simple waves, exact solutions, and limit and branch lines. In the final chapter the author expands his discussion of hodograph techniques, introduces the oblique shock in a perfect gas, and concludes the book with a stimulating discussion on the existence of smooth transonic flows. For the reader who is interested, there are some forty pages of notes and addenda which supply interesting historical footnotes to the text.

Certainly Mrs. von Mises and Professor Lundford are to be congratulated on the manner in which they completed the text, for the reader's immediate impression is that the book was written in its entirety by Professor von Mises. Of course, as in any book as detailed as the present one, it is always possible to disagree with certain features of the presentation. Thus, this reviewer felt at times a lack of Professor von Mises' apt physical description of fluid flow phenomena. This is particularly manifest in the presentation of characteristics and in the later discussion of simple waves. In addition, this reviewer did not always find a consistent level of treatment as evidenced, for example, by the unecessary restriction to a perfect gas in treating shock waves directly after introducing a general state equation. Some difficulty may also be encountered as a result of the manner in which some of the material is presented, an example of which is the introduction of one-dimensional shock structure before any consideration is given to the inviscid Hugoniot shock conditions.

In spite of these minor criticisms, to the reader who has some acquaintance with the field of compressible fluid flow and who desires a text which presents the more mathematical aspects of the subject, as well as to students and research workers more directly concerned with compressible flow as considered from a mathematical point of view, the book is highly recommended. In the years to come this book is certain to be recognized for its fund of information. It can be considered a fitting tribute to the author.

RONALD F. PROBSTEIN

Einige nichtlineare Probleme aus der Theorie der selbsttätigen Regelung. By A. I. Lurje. Akademie-Verlag, Berlin, 1957. xi + 167 pp. $3.60.

This monograph represents a revised and enlarged edition of papers written in the years 1945–1950. Its purpose is to bring the modern theory of nonlinear control to the engineers designing such control systems. Therefore, after careful exposition of methods and solutions, particular examples are treated in each chapter.

The book consists of four chapters: The canonical form of the equations in the theory of automatic control; the stability of control systems with one controlling element; self oscillations in control systems; behavior of a control system at the boundary of the stability region. The first chapter serves as a base for the following ones, which are independent of each other.

The state of the controlled system is described by n variables φ_k. In the direct control their deviation $\Delta\varphi_k$ from the desired equilibrium values determines the action ξ of a controlling element (e.g. motor) which should reestablish the equilibrium of the controlled system. In case of feedback a linear combination (called σ) of ξ, its derivatives and the $\Delta\varphi_k$ determines the action of the motor. The differential equations describing the state of the controlled system are linear; the nonlinearity enters through the characteristic of the controlling motor: $\dot{\xi} = f(\sigma)$. Various realistic functions $f(\sigma)$ are considered in the examples.

The canonical form of the differential equations describing the state of the controlled system is derived without the use of matrices, but with ample use of determinants and Dirac's delta function.

In Chapter II the stability in the large of the control system is investigated with as little restriction as possible concerning $f(\sigma)$. A Liapounoff function is constructed for the problem in question with the assumption that $\sigma f(\sigma) > 0$. The conditions for stability of a given system are determined by a system

of n quadratic equations, whose solution for $n > 2$ is tedious and for $n > 5$ extremely difficult (Lurje's own statement). In addition the special case $f(\sigma) = c\sigma + \Phi(\sigma)$ with $c = \text{const.} > 0$ and $\sigma\Phi(\sigma) > 0$ or $\sigma\Phi(\sigma) < 0$ is investigated.

In the third chapter Lurje studies the occurrence and stability of self oscillations in nonlinear systems by the method of Kryloff and Bogoliuboff and by the method of Poincaré. Both methods are explained in detail, based on a paper by Bulgakoff. Readers principally familiar with the methods will enjoy the discussion of their special merits and difficulties in the present problem.

The fourth and last chapter of the book is based on a paper by Bautin which in turn used a method suggested by Liapounoff in 1935. (Unfortunately no translation of this paper of Liapounoff seems to be available; his famous book, on "the general problem of stability" (1892) was translated into French; some chapters of the latter are expected to be known to the reader of Chapter II). The boundary of the stability region is considered to consist of dangerous and not dangerous portions. A disturbance starting outside, but close to the not dangerous portion will not deviate too much from a stable motion. On the other hand motions with initial values close to, but outside the dangerous boundary of the stability domain may deviate considerably from the stable motions.

The translation of Lurje's monograph into German is well done and certainly helps getting acquainted with these investigations, so important for the design of control systems.

I. Flügge-Lotz

Lectures on ordinary differential equations. By Witold Hurewicz. The Technology Press of the Mass. Institute of Technology, Cambridge, John Wiley & Sons, Inc., New York, and Chapman & Hall, Ltd., London, 1958. xvii + 122 pp. $5.00.

This book is a reprint of lecture notes of a course given by Hurewicz in 1943. It is a very clear and readable introduction to the fundamentals of the theory of systems of differential equations.

The first and second chapters deal with questions of existence and uniqueness of solution, the third deals with the properties of linear systems with constant or variable coefficients, the fourth with the theory of the singularities of second order systems, saddle points, nodes, foci, etc., leading up to a discussion of the periodic solutions of second order nonlinear differential equations.

For those who are interested in following the more modern developments, there is a list of twelve volumes dealing with various aspects of the theory of ordinary differential equations.

Richard Bellman

High-speed data processing. By C. C. Gotlieb and J. N. P. Hume. McGraw-Hill Book Co., Inc., New York, Toronto, London, 1958. xi + 338 pp. $9.50.

The field of high-speed data processing has become a highly specialised branch of the general field of computing, and there has been a need for an account of its peculiarities, equipment and techniques, although it is still in a state of rapid development. This book goes some way to fill this need, and is particularly suited to the newcomer to the computing field who expects to specialise in data processing, but those already experienced in the use of general-purpose computers will find many of their questions unanswered.

When trying to deal with a restricted aspect of the computing field which depends so much upon new and changing equipment there is a strong temptation to stray into the ramifications of computer design and programming principles and niceties, but some discussion of these aspects is unavoidable in the absence of books dealing with them specifically. Although such preliminaries occupy a little over half the book, an eye is kept on those aspects particularly applicable to data processing, and the effect of the special requirements of data processing of equipment compared with those of scientific computation are discussed in the earliest chapters, and throughout, actual machines are frequently referred to.

The treatment of representation of information concentrates upon alpha-numeric data and coded decimal systems. This considerably assists later discussion on programming and coding. Knowledge of the details of binary arithmetic is not essential and is correctly relegated to an appendix.

Modern data processing systems are generally centered around a 'general-purpose' type of computing and control unit which usually possesses an instruction code specially adapted to the class of work to be carried out; thus there is special emphasis on input, output, store transfers and comparisons and so on. The organization within the central unit depends upon well established principles, and the discussion of the details of the central unit and various storage systems would seem to be excessive compared with the treatment of the various means of input and output upon which data processing so largely depends.

The discussion of instruction code types and address systems is followed by a lengthy section on programming and coding, and the authors have adopted a hypothetical instruction code as a model. This code is readily understood and remembered, being single-address coded-decimal with mnemonic letter code for operation type. It contains special features to assist coding for data processing, including block transfers, input and output buffers, a number of operations to assist double length operations within the accumulator and storage of discriminations. It does however have the disadvantage, for the beginner, of packing two instructions to a word; these are addressable only in pairs, and this complicates the organisation of control transfers and requires the use of redundant 'skip' instructions. It has to be admitted that some machines are just like this, and in the words of the authors; 'This awkward feature of the Hypothetical Machine has been retained to allow the programmer his occupational prerogative of complaining about the instruction code'.

The main techniques of programming are illustrated via a number of simple problems of data processing type and occupies some 60 pages. The use of autocodes of various kinds is becoming common, and this method of coding is described only in the final chapter.

A considerable simplification in the sections on coding would have been achieved, without lack of reality, if this tendency had been accepted, and coding been described in terms of even a simple autocode. The use of symbolic addresses alone would have been of help, to the beginner. The treatment of machine organization could have been shortened and the more interesting features of data reduction dealt with in the later sections dealing with the use of files, sorting, selecting and practical applications, could have received closer treatment, as for instance, on discussion of the various ways of marking file and record endings, the use of additional tracks, and forward and backward reading.

Examples of applications have been taken from the commercial fields including insurance, accounting, planning and scheduling, and there is some discussion of the use of processors as simulators. This latter use would seem to have very great potential for the future development of data processing and would have justified expansion. The authors have in fact kept their feet well on the ground in discussing data processing as it now exists, and there is little reference to its future. Among aspects of the future of data processing which one would like to see discussed are the implications of the treatment of the results of continuously recording scientific experiments, an expanding field of data processing which calls for high processing speeds; the possible effects of increased processing speeds on code systems; the likely effects of new electronic techniques; and so on.

The extension of data processing to wider fields is still largely dependent upon the elimination of the human being at the early stages, and the presentation of large amounts of treated information in readily assimilable forms. Future developments would seem to depend upon the design of special input and output devices for translating original data, as occur in banking and national census and market research, into directly usable media without human intervention; and high speed multi-curve plotters for ready assimilation by the reader.

The book makes a good introduction to the current problems and techniques of data processing, it is excellently clear in style, and contains, not least in importance, an extensive bibliography.

T. Pearcey

An introduction to combinatorial analysis. By John Riordan. John Wiley & Sons, Inc., New York, and Chapman & Hall, Ltd., London, 1958. x 244 pp. $8.50

The author interprets combinatorial as "anything enumerative" and presents a fairly systematic survey of the subject with particular emphasis on the developments of the last fifty years and on the use of generating functions. The book should be of interest not only to specialists in the field but should serve as a useful reference as well since, between the text and the problems (of which there are about 200), the book contains the solutions to a considerable number of problems. In addition to the classic

material on permutations, combinations, etc., there is a chapter entitled "Partitions, Compositions, Trees, and Networks" plus two on "Permutations with Restricted Position" much of the material in which is either new or of recent origin.

Although the reviewer could find no specific criticism to make of the book, he did not find it very inspiring. This may be more a consequence of the reviewer's attitude toward the subject than the authors presentation although the authors style of writing does not seem to convey much enthusiasm.

G. F. NEWELL

Electronic digital computers. By Franz L. Alt. Academic Press Inc., New York and London, 1958. x 336 pp. $10.00

The book is divided into five parts: 1. Introduction; 2. Automatic digital computers; 3. Coding and programming; 4. Problem analysis; 5. Matching problems and machines. Parts 1 to 3 are standard material but 4 and 5, which make up well over half of the book, present matter not previously collected in book form. In part 4, methods of numerical analysis are classified and discussed in relation to their usefulness and efficiency when used with computers; and in part 5 the impact of computers on scientific and engineering research is demonstrated by examples drawn from many fields. The exposition is clear and concise, and geared to a mathematical level far below that of the other books in this series.

WALTER F. FREIBERGER

Ordinary differential equations. By Wilfred Kaplan. Addison-Wesley Publishing Co., Inc., Reading, Mass., 1958. xv 534. $8.50.

This is an excellent introductory text in the field of differential equations. It is carefully written and carefully planned, with many important illustrative examples, illuminating graphs, and a large number of exercises.

The first part, about three hundred pages, is devoted to a thorough study of linear differential equations with constant coefficients, n-th order equations, and linear systems. For this latter purpose, matrices are introduced and applied. A number of engineering applications are given and there is a very helpful discussion of the interconnection between various engineering terms and mathematical vocabulary. A long chapter, fifty pages, is devoted to the fundamental technique of power series solutions, and a brief chapter, ten pages, perhaps too brief, to numerical solution via difference techniques.

The penultimate chapter contains a very readable introduction to the study of periodic solutions of nonlinear second order differential equations via phase plane analysis. In the final chapter, we find existence, uniqueness, and convergence theorems which the author has wisely postponed to the end of the volume.

The book is heartily recommended for college classes and for those who wish to prepare themselves for the more advanced theory of differential equations and its applications.

RICHARD BELLMAN

Boundary layer research. Edited by H. Görtler. Springer-Verlag, Berlin, Gottingen, and Heidelberg, 1958. xii 411 pp. $16.20.

This book differs notably from the usual symposium proceedings (so often rather aimless and full of ticket papers). A symposium was actually achieved, by felicitous choice of a subject of neither too wide, nor too narrow, a scope. And while geography played clearly a role in the selection of the participants, the remaining freedom of choice was employed to ensure an average level of contributions well above that usual at such international meetings.

The 31 papers and 22 notes range over the whole field of boundary layer research, with the keenest interest discernible, perhaps, in the subjects of stability and separation.

The volume offers more than a formal record of proceedings, due to the inclusion of a part of the impromptu discussions. These make the meeting come partly alive again, and it was reluctance to

miss any of them (and especially the pungent and illuminating remarks offered by some of the English participants), almost as much as reluctance to miss any of the first-class papers, which made this reviewer read through all 411 pages. It must have been fun to be at the meeting, and it is fun to read its record. Apart from which, the book will be needed on the shelves of all those with a more than transitory or platonic interest in boundary layers.

R. E. MEYER

Elementary statistical physics. By C. Kittel. John Wiley & Sons, Inc., New York, and Chapman & Hall Ltd., London, 1958. ix 228 pp. $8.00.

This book is based upon a series of lectures given to beginning graduate students in physics and is written in a style quite similar to the author's earlier book "Introduction to Solid State Physics". Unlike many authors who have tried to write a book that will serve as both a text and a reference but succeed in doing neither, this book is only an elementary text. For the more difficult topics such as ergodic theory, phase transitions, etc., frequently described in books of similar titles, the author simply refers the reader to the more advanced works. The scope, however, is very broad and includes besides the usual statistical mechanics, short sections on irreversible thermodynamics, Brownian motion, noise theory and other topics not always found in an elementary book.

Although the author sacrifices rigor at times, both in the mathematics and in the physics, to preserve simplicity, this seems to be unavoidable in a book of this type. At no place does the author embark on long tedious calculations or deviate very much from a fairly constant level of difficulty. Particularly because of this, the reviewer believes that this is the best elementary text book presently available on this subject. The book is primarily for physicists, however.

G. F. NEWELL

An introduction to multivariate statistical analysis. By T. W. Anderson. John Wiley & Sons, Inc., and Chapman & Hall Ltd., London, 1958. xii 374 pp. $12.50.

This textbook gives a very systematic treatment of multivariate analysis. Throughout most of the book the methods of analysis are derived by the maximum likelihood method (for estimation) and the likelihood ratio criterion (for testing hypotheses). The first of these two methods is known to have certain optimality properties; the second one also has some desirable large sample features and often leads to reasonable tests. In this way the author has organized the material successfully and presents a unified treatment of the many different subjects that belong to multivariate analysis.

After introducing the reader to the multivariate distribution the author describes how to estimate the mean vector and the covariance matrix and how the empirical correlation coefficients are distributed. One chapter is devoted to the T^2-statistic and the Behrens-Fisher problem.

The next chapter deals with a different type of problem, namely the classification of observations into one or two or more normal populations. This topic is discussed from the point of view of statistical decision functions in terms of Bayes procedures, admissible classes, etc.

After a detailed discussion of testing various multivariate hypotheses the author turns to principal components and canonical correlations, which are treated in a very lucid way.

In the last chapter some more advanced problems are sketched briefly, and it is only to be regretted that they were not given more space in the book; this is true especially about factor analysis.

An appendix on matrix theory ends the book. At the end of each chapter a number of problems are given.

This book is an attractive and representative example of the modern point of view on theoretical questions among mathematical statisticians in this country. It should be very useful to anybody interested in multivariate analysis.

ULF GRENANDER

Elasticity and plasticity. By J. N. Goodier and P. G. Hodge, Jr. John Wiley & Sons, Inc., New York, and Chapman & Hall, Ltd., London, 1958. ix + 152 pp. $6.25.

This book is the first volume of a series of surveys in Applied Mathematics. In the preface the authors make a modest statement, saying that their work would be devoted only to such a group of problems in the domain of the theory of elasticity and plasticity which were dealt with in publications rather inaccessible or published in non-familiar languages (non-familiar from the English speaking people's point of view). So it was not the authors' aim to present an exhaustive and proportioned survey of all the branches of the subject in question, but rather to focus attention on those important latest achievements which, according to their opinion, are little known to the readers whose first language is English. What is known and easily accessible has been omitted or only touched upon. Nevertheless, the book plays a much more important role because it represents a kind of monographical approach to some fields of the theory of elasticity and plasticity.

The approach to the subject in Part I (The Mathematical Theory of Elasticity, by J. N. Goodier) is different from that in Part II (The Mathematical Theory of Plasticity, by P. G. Hodge, Jr.), both in essence and form. However, each of them has its own merits, the first part being more a survey, the second a survey with an attempt towards monographical approach.

Part I (47 pages) deals successively with: two-dimensional problems, the problems of holes and fillets (without and with reinforcement), mixed boundary value problems, anisotropic elasticity, thermal problems, three-dimensional contact problems, wave propagation, seismical and vibrational problems. Bibliography contains 128 entries (65 Western, 63 Soviet).

Part II (96 pages) deals successively with the foundations of the theory of perfectly plastic bodies, of strain-hardening bodies, piecewise linear plasticity, the minimum principles, and a number of applications of the theory, such as bending of planes and shells, plane strain and plain stress, beams, rods, and miscellaneous problems. A separate chapter is devoted to Soviet, Polish, Hungarian, and Chinese papers. Bibliography contains: 345 entries (182 Western, 149 Soviet, 11 Polish, 2 Hungarian, 1 Chinese).

The closing part of the book includes Author Index and Subject Index (8 pages).

The book is interesting and refreshing. Both authors, authorities in their specialities, made a worthy contribution, which is very useful and serves the purpose: for experts, the book represents a valuable contemporary informative source (with a detailed list of references); for non-specialists, it may be considered as an easily understandable introduction to the subject.

W. Olszak

Games and decisions. By R. Duncan Luce and Howard Raiffa. John Wiley & Sons, Inc., New York, and Chapman & Hall, Ltd., London, 1957. xix + 509 pp. $8.75.

Mathematical fields, as all other subject areas, are in constant need of redefinition. Applied mathematics is no exception to this rule. One of the most interesting phenomena of the past couple of decades is the wide application of mathematical techniques to areas that previously lacked them. The historian of the future may well regard the present era as one of the times when mathematics was most healthily challenged and reinvigorated by the posing of new problems, arising from attempts to apply mathematics to new fields of study.

Among the new areas of application are the nonphysical (or behavioral) sciences such as anthropology, biology, economics, genetics, psychology, management science, and sociology. The book under review gives a wide-ranging discussion of a number of these new kinds of applications, and may properly be regarded as a book on applied mathematics. This is true even though the traditional tools of applied mathematics (such as differential and integral equations) nowhere appear in the book. Instead, the mathematics of convex sets and other kinds of mathematics formerly in the "pure" category are used to state and criticize the results of game theory.

Game theory was put on its mathematical feet in the epic paper of John von Neumann in 1928. This paper was almost unnoticed for 16 years until, in 1944, the treatise *Theory of Games and Economic Behavior*, by von Neumann and the economist Oskar Morgenstern appeared in 1944. There was a brief period of further dormancy of the theory until about 1948 when the first of a flood of papers started appearing on the subject. To show how vast the subject has become in this short time, a bibliography of articles in game theory up to 1957, recently compiled by the reviewer and his wife, lists more than

1000 papers and books on the subject during that period. Some of these articles are expository in nature and were written to introduce the subject to a new audience. However, a substantial fraction of these papers contain the meat of a new result.

It is manifestly impossible for a research mathematician, say, to attempt the reading of this many papers unless he has a very serious professional interest in the subject. But it is completely impossible for a non-mathematician, say an interested behaviorial scientist, to plow through all that material and still maintain himself in his own profession. The authors (both of whom received their principal training in mathematics) have undertaken the task of surveying the field. Because many of the people who want to know these results are not technically trained mathematicians, the authors have merely included the statements and critique of the results together with a bibliography of references for those who wish to probe deeper into the mathematical background.

Although both authors are mathematicians, Luce has concerned himself very much with sociology and Raiffa has made extensive studies of statistics. Both these biases appear in the book. For instance, we find that they talk of a player's "security level" instead of his mathematical expectation—clearly sociological terms. Also the very complete discussion of the statistical decision problem is a reflection of Raiffa's special interests. These emphases are to the good, since they reflect the parts of the theory on which each of the authors has concentrated most fully. It might be noted that the discussion of continuous games is fairly sketchy.

A brief summary of the contents of the book is that it covers the two-person zero-sum game, several of the current solution theories of the n-person non-zero sum game, individual decision making (statistical decisions), and group decision-making (the welfare problem). The content areas to which these subjects are most closely related are sociology, statistics, economics, and management science.

The authors are to be commended for their courage in undertaking such a formidable task as the survey of such a large and recent area of study. The result of their labors is a marvelous exposition of most of the theory. This book is certain to be influential in the future training of students working in the theory as well as in stimulating work on specific research problems. It should be of interest to anyone who wants to find out about new kinds of applied mathematics from either a cultural or technical point of view.

The decision as to whether or not a behavioral scientist or applied mathematician should have this book on his bookshelves is easy—he should.

G. L. Thompson

An introduction to fluid dynamics. By G. Temple. Oxford University Press, New York, 1958. xi + 195 pp. $4.00.

The preface states that "the object of this book is to provide an introduction to fluid dynamics, primarily for students reading for honors in mathematics and theoretical physics." One cannot but approve the object and simultaneously acknowledge that the author has achieved it with brilliant success.

Keeping within the bounds of inviscid continuous fluid and with an eye firmly fixed on the physical problem, Temple has produced in 190 pages a fascinating account of hydrodynamics. Particular emphasis has been placed on the "fluid body" that is a portion of fluid which always consists of the same fluid particles. In this connection the argument of 1-3 concerning the application of the laws of motion appears to be unconvincing, for it starts from the tacit assumption that the internal forces form a self-equilibrating system.

The reader is led by easy stages through elementary notions to sources, doublets and vortices, to distributions of these singularities and the action on a body in a uniform stream. Here, in deriving the Kutta Joukowski theorem the author obtains a drag term due to the total source strength in Green's equivalent stratum. To the reviewer this term seems to be necessarily zero, for the normal velocity on the surface of a body at rest in a uniform stream vanishes and it is to this normal velocity that the stratum is due. The text then goes on to conformal mapping, free streamlines, design of wing profiles, axisymmetric flow, and finally slender body theory applied to solids of revolution and checked by exact solutions for the ovary ellipsoid.

The book can be heartily recommended.

L. M. Milne-Thomson

Mathematics of physics and modern engineering. By I. S. Sokolnikoff and R. M. Redheffer. McGraw-Hill Book Company, Inc., New York, Toronto, London, 1958. ix + 810 pp. $9.50.

The present volume shares with its well-known predecessor, *Higher Mathematics for Engineers and Physicists*, the aim of providing a discussion of those topics beyond the calculus which are of importance in engineering and physics. Thus, mathematical concepts are first presented in a precise manner and are then illustrated in many cases by examples and problems drawn from engineering and physics, and throughout the book the authors have attempted to preserve this balance between the formal aspects of the subject on the one hand and their application to physical problems on the other. The text is divided into nine chapters which deal with ordinary differential equations, infinite series, functions of several variables, vectors and matrices, vector field theory, partial differential equations, complex variable, probability theory, and numerical analysis, with short appendices on determinants, the Laplace transform, and the Riemann and Lebesgue integrals. Since the chapters are largely independent, the book can easily be adapted to varying teaching requirements and will also prove useful for reference purposes.

W. H. Reid

Introduction a L'Algèbre Supérieure et au Calcul Numérique Algébrique. By L. Derwidue. Masson et Cie, Paris, 1957. 431 pp. $15.80.

The selection of topics in this text book on advanced algebra is especially happy from the point of view of engineers and physicists. The author notes in his preface that, while good texts covering these topics are available in English and German, the literature in French is very sparse. The principal subjects covered are: systems of linear equations and determinants, polynomials, matrices and matrix eigenvalue problems and the stability criteria of Routh, Hurwitz and Schur. In all of these topics great emphasis is placed on bringing the theoretical results down to a form which permits numerical calculation. It is assumed that the reader has access to a desk calculator. Many sample calculations are given indicating appropriate layouts for recording intermediate results and pointing out methods for carrying along "running checks." Ill-conditioned systems are mentioned and the advantages of selecting the largest divisor in elimination methods are demonstrated but the general round-off problem is not touched. High speed automatic digital computers are not discussed. In the final chapter there is a short introduction to abstract algebra: groups, rings, fields, etc.

All in all, the author has succeeded in his attempt to provide a transition between elementary algebra and the specialized fields of numerical analysis and modern algebra.

S. H. Crandall

Space-charge waves and slow electromagnetic waves. By A. H. W. Beck. Pergamon Press, New York, London, Paris, Los Angeles, 1958. xi + 396 pp. $15.00.

This book is the eighth volume in the International Series of Monographs on Electronics and Instrumentation. It covers a great deal of ground and on the whole the material is well presented, with more or less complete reference to pertinent published papers. There is a fair amount of mathematics used, although this entails not much more than a knowledge of the simpler properties of Bessel functions and some matrix algebra. There are some places where the logical structure is rather poor and this is complicated further by numerous misprints in some of the more mathematical sections. Also, the author occasionally quotes results from published papers in a highly condensed form without complete definitions of symbols used. This is again complicated by misprints and it was not always easy for the reviewer to follow the deductions without reference to the paper in question. However, these difficulties do not detract from the general worth of the book and the reviewer believes that the applied mathematician who wishes to become thoroughly acquainted with the theory underlying the various microwave valves in use today may do so by reading this book which is obviously written by one with an extensive knowledge of the subject. A condensed summary of the contents follows:

A short general introduction is given in Chapter I in which various types of amplifiers are briefly mentioned. Chapter II gives a condensed account of Maxwell's electromagnetic theory. Propagation modes are discussed for rectangular and circular waveguides and expressions for the power flows are obtained. Vector and scalar potentials and the Hertzian vectors are introduced. Chapter III deals with various slow wave structures and in particular with disk-loaded waveguides, interdigital delay lines and both sheath and tape helices. Their properties are discussed in relation to the Brillouin diagram (frequency versus phase constant). Chapter IV is devoted to space-charge wave theory and topics include the determination of the space-charge reduction factors for cylindrical and annular beams in cylindrical tunnels; space-charge waves on Brillouin beams and on confined beams; space-charge waves in accelerating fields and in crossed fields; the distribution function approach for multivelocity electron beams and plasma oscillations. Chapter V utilizes some of these results in discussing the matching of input conditions and space-charge waves on cylindrical, annular, Brillouin and confined beams, and on beams in helices.

Chapter VI is concerned with space-charge waves in klystrons while Chapter VII is devoted to travelling-wave tubes and backward-wave oscillators. Some crossed field devices are discussed in Chapter VIII while some special space-charge wave devices, such as the two-beam tube and the transverse current T.W.A. (travelling-wave amplifier), are mentioned in Chapter IX. The final Chapter deals with noise phenomena in space-charge wave devices and after discussing shot noise in diodes, the noise factor for valves with input resonators and the noise figure of T.W.A.'s, the general theory of noise in beams, based on the theorems of power flow in space-charge waves, is presented in the matrix formulation adopted by Haus and Robinson, although the proof of the noise invariants is omitted. The smoothing of current fluctuations near the potential minimum is also briefly discussed. There are twelve Appendices, mostly mathematical in content, and these are followed by a few questions on each chapter, a list of recent references, a list of the major symbols used throughout the book and, finally, the index.

J. A. Morrison

Selected papers on quantum electrodynamics. Edited by Julian Schwinger. Dover Publications, Inc., New York, 1958. xvii + 424 pp. $2.45.

This book is a collection of thirty-four papers taken from the literature of quantum electrodynamics—papers which together with a preface summarize the present state of the theory. In summing up the problems still facing quantum electrodynamics the editor points out that "the real significance of the work of the past decade lies in the recognition of the ultimate problems facing quantum electrodynamics."

Papers are included in this collection by the following: Dirac, Fermi, Fock, Poldolsky, Jordan, Wigner, Heisenberg, Weisskopf, Block, Nordsieck, Foley, Kusch, Lamb, Retherford, Bethe, Schwinger, Oppenhemier, Tomonga, Pauli, Villars, Feynman, Dyson, Karplus, Klein, Källen, and Kroll.

Of these papers twenty-nine are in English, three in German, and one each in French and Italian.

R. Truell

Computability and unsolvability. By Martin Davis. McGraw-Hill Book Company, Inc., New York, Toronto, London, 1958. xxv + 210 pp. $7.50.

This book is an introduction to the theory of computability and noncomputability, usually referred to as the theory of recursive functions, a branch of pure mathematics. In the light of recent developments in computers, decision problems, i.e., problems "which inquire as to the existence of an algorithm for deciding the truth or falsity of a class of statements," are potentially of interest to other than pure mathematicians. The work is an outgrowth of a graduate course at the University of Illinois and series of lectures given at the University of Illinois and the Bell Telephone Laboratories.

The major part of the book is self-contained and assumes no particular mathematical training on the part of the reader. A degree of mathematical maturity, in particular the ability to follow abstract proofs, is, however, necessary. Acquaintance with elementary mathematical logic is desirable.

The concept of Turing machine is made central to the development. Thus Turing's approach is combined with the methods of Gödel and Kleene to present the various aspects of the theory of computability in a unified manner. The general theory is applied to combinatorial problems, problems related to Hilbert's Tenth Problem, and systems of symbolic logic. The general theory is extended to include the Kleene hierarchy of arithmetical predicates, computable functionals, and the classifications of unsolvable decision problems.

A. A. GRAU

Modern geometrical optics. By Max Herzberger. Interscience Publishers, Inc., New York, London, 1958. xii + 504 pp. $15.00.

This work is both a treatise and a text on the methods of geometrical optics. As far as the reviewer is aware it is the only book of this kind in English (other than the notes of Lumeberg which were never published).

There are essentially two distinct sections of the book. The first part is concerned with the ray tracing and the calculation of optical systems; this consists of about the first thirteen chapters. Much of the second section of the book is concerned with the general laws of geometrical optics. The characteristic function methods of Hamilton, the general laws of image formation, the properties of concentric systems, and the properties of rotation symmetric systems are naturally an important part of the book.

The contents of the book are listed as follows:

Part I, Ray Tracing; Part II, Precalculation of Optical Systems; Part III, General Laws; Part IV, Concentric Systems; Part V, Rotation-Symmetric Systems; Part VI, Approximation Theory for Normal Systems; Part VII, Third and Fifth Order Image-Error Theory; Part VIII, Interpolation Theory of the Optical Image; Part IX, Optics in General Media; Part X, Appendix.

The only topics that the reviewer felt were noticeably missing, topics that might have been included, are those of integral invariants and perhaps something about electron optics; it may be that this is asking for too much.

The book represents a tremendous amount of work. It is well written, and it should be the standard work in this field for a long time.

R. TRUELL

Principles of quantum electrodynamics. Translated from the German by J. Bernstein, with additions and corrections by Walter E. Thirring. Academic Press Inc., New York, London, 1958. xv + 234 pp. $8.00.

This book is a text on the quantum electrodynamic part of field theory. It is more advanced in nature than the book of Wentzel entitled "Quantum Theory of Fields." The text is the result of an effort by the author to present mainly what cannot be found elsewhere in books on this subject. There are four parts of the text divided as follows:

I. General Introduction: Units and Orders of Magnitude; Classical Electrodynamics; General Formalism of Quantum Theory of Fields. II. Free Fields: General Discussion; Special Fields; Matrix Elements; Fluctuation Phenomena. III. Fields With External Sources: General Formulae; Emission of Light; The Dirac Field in an External Electric Field; The Limitations of Measurability. IV. Interacting Fields: General Orientation; Scattering Processes; Renormalization Theory; Higher Order Correction; Outlook.

Appendix I, Dirac Matrices. Appendix II, Green's Functions (Relativistic Wave Equation)

While this text is intended for the advanced student, there are parts such as section 16 of Part IV which will be of interest even to those who only want to know what the problems are at the present time.

R. TRUELL

Theorie schallnaher Strömungen, By K. G. Guderley. Springer-Verlag, Berlin, Gottingen, Heidelberg, 1957. xv + 376 pp. $10.07.

The study of transonic flow as a special field was stimulated by the failure of classical treatments of compressible flow, both by theoretical and experimental means, to describe flow phenomena in the Mach number range close to critical conditions. This gap arises because, near Mach number one, the linearized theory of compressible flow is invalid and normal High Speed Wind Tunnels become choked.

Much of the fundamental theory of transonic flow was developed by Dr. Guderley himself and it is appropriate that he should now present a connected account of his own and other contributions to the subject. Although his book is primarily concerned with transonic small disturbance theory based on Tricomi's equation it is self contained and the theory is carefully introduced with the necessary background of fundamental Gas Dynamics. The book will be of considerable interest to Applied Mathematicians and Theoretical Aerodynamicists. A great deal of space is given to special solutions of Tricomi's equation and their use in solving the problem of flow past a double wedge profile in a sonic stream. This is a field to which the author devoted many years of research, previously published in somewhat inaccessible form. It is certainly valuable to have this work in a single volume. The book would, in the reviewer's opinion, have been better balanced if more space had been given to other approaches to transonic small disturbance theory, notably by Oswatitsch, and to a fuller description of the purely numerical solution of transonic flow problems. The book is welcome as the first, and much needed, book on transonic flow and is strongly recommended to Fluid Dynamicists. There is still room, however, for a more complete account of the theory and also for an account of the difficult experimental work in this field. Further, since the book was written a number of transonic flow problems have been solved in the U. S. S. R. by Chushkin, using the successful numerical technique of Dorodnitsin for second order partial differential equations of mixed type. This approach certainly competes with the classical treatment based on Tricomi's equation. It is hoped that future editions of Dr. Guderley's book will include an account of this recent work.

There are 11 chapters in the book. Following a first chapter devoted to the basic principles of Steady Gas Dynamics the transonic similarity concept is introduced in Chapter II. The third chapter contains a brief account of Linearized Transonic Flow. In Chapter IV exact solutions of the simplified transonic potential equation are given for flow through a Laval nozzle and flow of a sonic jet. Chapter V gives a fairly complete account of the hodograph method leading to Tricomi's equation when transonic similarity is introduced. Chapter VI is largely a physical discussion of transonic flow phenomena in the hodograph plane. The author's special solutions of Tricomi's equation based on the singularity at the sonic point are described in full in Chapter VII. This is followed, naturally, by the solution of problems of flow with Mach number 1 in Chapter VIII. Chapter IX presents methods for determining flow fields with free stream Mach numbers slightly different from 1, based essentially on a linear perturbation of the solution given in Chapter VIII. In Chapter X some special points are discussed such as the reflection of disturbances at a sonic line. The final chapter is concerned with the extension of Guderley's basic theory to axially symmetrical flow.

M. Holt

Advances in applied mechanics. Editors: H. L. Dryden and Th. von Kármán. Academic Press, Inc., New York, 1958. x + 459 pp. $12.00.

The appearance of a new volume in this series, with its outstanding surveys of selected topics in Applied Mechanics, is always most welcome. For it is only by means of review articles of the type presented here that one can even attempt to keep abreast of developments in so inclusive a field as that of Applied Mechanics, an inclusiveness well demonstrated by the titles of the present articles: "Supersonic air ejectors" by J. Fabri and R. Siestrunck, "Unsteady airfoil theory" by A. I. Van De Vooren, "The theory of distributions" by Charles Saltzer, "Stress wave propagation in rods and beams" by H. N. Abramson, H. J. Plass and E. A. Ripperger, "Problems in hydromagnetics" by Edward A. Frieman and Russell M. Kulsrud, "Mechanics of granular media" by H. Deresiewicz, and "Condensation in in supersonic and hypersonic wind tunnels" by P. P. Wegener and L. M. Mack.

W. H. Reid

Notes on analog-digital conversion techniques. Edited by Alfred K. Susskind. The Technology Press of Mass. Institute of Technology, Cambridge, and John Wiley & Sons, Inc., New York, 1957. x + 877 pp. $10.00.

The articles, by various authors, in this book had their origin in a special summer program on the subject at M. I. T. in 1957. The subject matter is divided into three parts: systems analysis, engineering analysis of devices and a case study, of which the first bears most interest to mathematicians, in particular Chapter 2 on the theory of sampling and the theory of quantizing, and Chapter 3 on codes.

W. F. FREIBERGER

Elementary mathematical programming. By Robert W. Metzger. John Wiley & Sons, Inc., New York, and Chapman & Hall, Ltd., London, 1958. ix + 246 pp. $5.95.

This book is addressed to industrial engineers and others with little mathematical background who wish to know the mechanics of programming without being burdened with proofs or conceptual discussions. It is lucid and provides many practical illustrations but does not discuss the limitations of the very simplified models.

WALTER FREIBERGER

Notes on Operations Research 1959. Assembled by the Operations Research Center, M.I.T. The Technology Press, Massachusetts Institute of Technology, Cambridge, Mass., 1959. viii + 256 pp. $4.00.

These are the lecture notes for a special program in operations research that members of the Operations Research Center at M.I.T. conducted for persons from European NATO countries at the Center for Experimental Aerodynamics in Brussels in August 1959. The scope of the book is indicated by the following table of contents: Introduction (P. M. Morse)—Probability (G. P. Wadsworth)—Search (B. O. Koopman)—Markov Processes (P. M. Morse)—Queuing (H. P. Galliher)—Control Processes (R. A. Howard)—Organization of Operations Research Groups (P. M. Morse)—Sequential Decision Processes (G. E. Kimball and R. A. Howard)—Reliability and Maintenance (G. E. Kimball)—Information theory (B. O. Koopman and G. E. Kimball)—Production Scheduling (H. P. Galliher)—Simulation of Random Processes (H. P. Galliher)—Bibliography.

The measurement of power spectra. By R. B. Blackman and J. W. Tukey. Dover Publications, Inc., New York, 1959. x + 190 pp. $1.85.

This is a republication of "Power Spectra from the Point of View of Communications Engineering", which originally appeared in Volume 37 of the Bell System Technical Journal.

Les Principes de la Théorie Electromagnétique et de la Relativite. By Marie-Antoinette Tonnelat. Masson et Cie, Editeurs, Paris VI^e, 1959. 394 pp. $10.25.

This book is concerned with the study of the principles of classical and relativistic electromagnetic field theory and gravitational field theory. The discussion of the principles of electromagnetic theory includes among other things, discussion of the electron theory of Lorentz and the electrodynamics of bodies in motion. The fifth chapter begins the discussion of relativity and the remaining three quarters of the book does an excellent job of this. The only feature of the book that detracts from its value is the rather poor printing job, especially the tensor indices which are sometimes illegible.

ROHN TRUELL

Theory of functionals and of integral and integro-differential equations. By Vito Volterra. Dover Publications, New York, 1959. xiv + 226 pp. $1.75.

This is an unabridged republication of the English translation of Volterra's work published by Blackie & Son Ltd. in 1930. A preface by Professor G. C. Evans and a biography and bibliography by Sir Edmund Whittaker (originally published as an obituary note) have been added.

Plastic analysis of structures. By Philip G. Hodge, Jr. McGraw-Hill Book Co., Inc., New York, Toronto, London, 1959. xiv + 364 pp. $10.50

The point of view that the rational analysis and design of structures composed of elastic-ductile materials requires inclusion of plastic action in the theory of structures has been attracting more interest recently, as evidenced by the increasing number of books published on the subject. The present book constitutes a systematic and up-to-date account of the theory of plastic analysis of structures and it is broader in coverage than its predecessors. Structures treated in the book include beams and frames in bending, beams under combined stresses, plates and shells and slabs with cut-outs. The subject matter is approached from the view-point of limit analysis. Emphasized are two basic theorems which enable lower and upper bounds on the load carrying capacity of structures to be determined. Part I, *Bending of Beams and Frames*, treats the application of plastic methods to frame-type structures. It also deals with such topics as elastic plastic deformations, variable and repeated loading and procedures of design for economy in materials. Part II, *Structures Under Combined Stresses*, is primarily concerned with combined stresses in beams, circular plates, and circular cylinderical shells. More general plate and shell problems and problems in plane stress have also been covered. Finally the book closes with a brief introduction to some of the problems encountered in the dynamic loading of plastic structures. References and Problems are given at the end of each chapter.

The clear presentation and reasonably elementary level of mathematical tools employed will undoubtedly make this text attractive to the interested students (both graduate and undergraduate) and also to the practicing engineer. However the structural engineer is probably too much elasticity-minded and may need additional convincing before accepting such concepts of plasticity theory as rigid-plastic action, plastic hinges and plastic collapse. The critical presentation and discussion of some experimental work, in addition to the test results given in the text, may prove helpful in this respect.

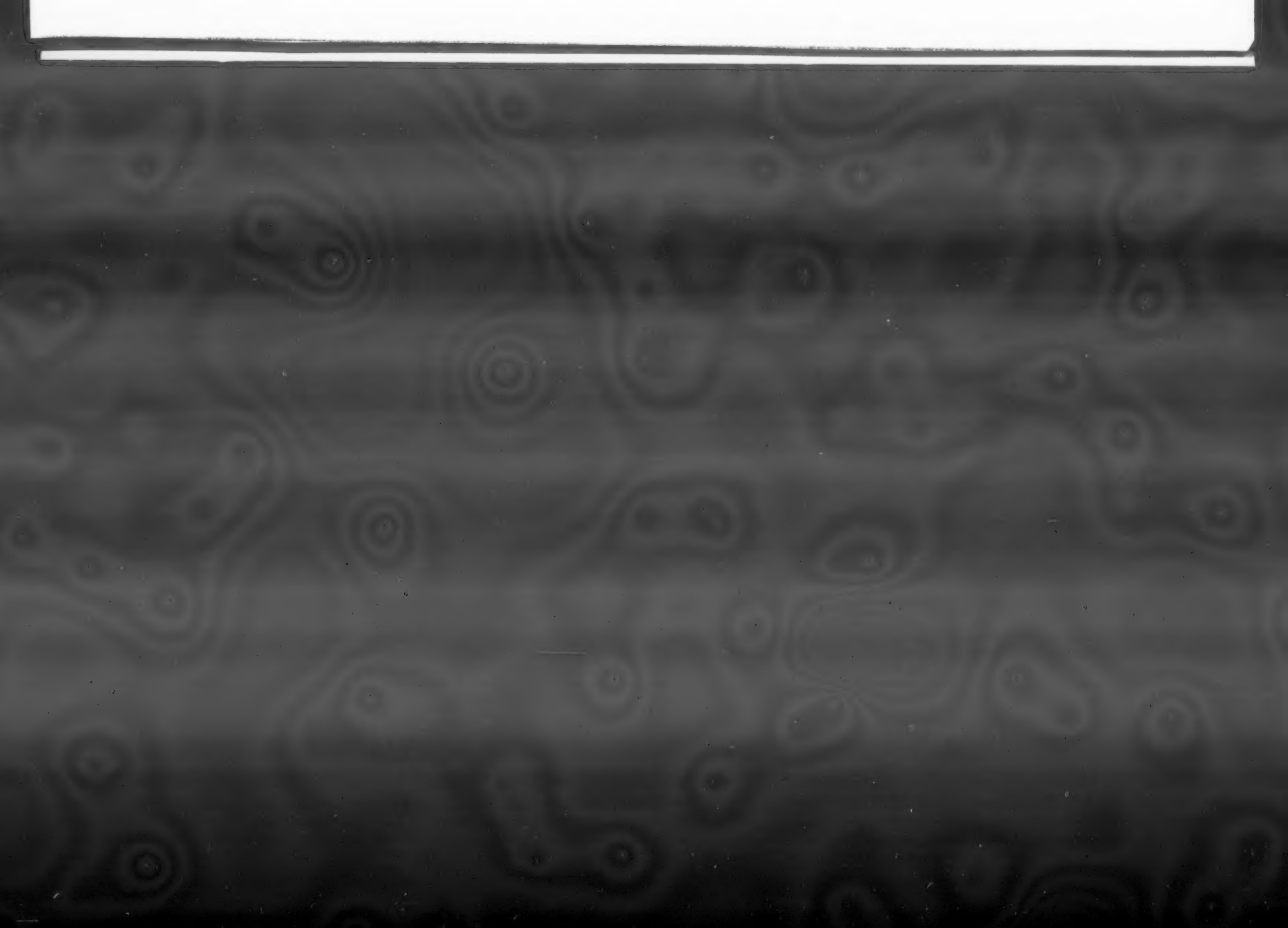
It may also be noted that questions concerning the effects of geometry changes and strain-hardening do not receive extensive attention in the book. The question of plastic stability is hardly considered in the book and beneficial effects of geometry changes and strain-hardening on the load carrying capacity of structures are mentioned only briefly. Some systematic discussion of these topics in the future editions of the book would increase the value of this interesting work.

E. T. Onat

Notice of

SYMPOSIUM ON PLASTICITY

The Second Symposium on Naval Structural Mechanics will be held April 5, 6, and 7, 1960 at Brown University, Providence, R. I., under the joint sponsorship of the Office of Naval Research, Department of the Navy, and of Brown University. The Symposium will be devoted exclusively to the field of plasticity. The program will consist of critical surveys in selected areas and of reports on original research, with ample time for discussion. The organizing committee consists of Professors E. H. Lee and P. S. Symonds (co-chairmen), D. C. Drucker and W. Prager.



CONTENTS

P. M. Chirlian: Restrictions imposed upon the unit step response of linear phase shift networks . . . 225
R. Bellman and J. M. Richardson: On the application of dynamic programming to a class of implicit variational problems . . . 231
M. Lotkin: Determination of characteristic values . . . 237
A. H. Zemanian: Some properties of rational transfer functions and their Laplace transformations . . . 245
H. R. Bailey: Heat conduction from a cylindrical source with increasing radius . . . 255
J. Meixner and C. P. Wells: Improving the convergence in an expansion of spheroidal wave functions . . . 263
W. Freiberger and U. Grenander: Approximate distributions of noise power measurements . . . 271
S. C. R. Dennis, A. McD. Mercer and G. Poots: Forced heat convection in laminar flow through rectangular ducts . . . 285
C. A. Hollingsworth: [illegible] . . . 297
W. T. Kyner: On a free boundary value problem for the heat equation . . . 305

Notes:

H. S. Tan: On a special Bolza variational problem, and the minimization of supersonic [illegible] . . . 311
G. A. [illegible]: Note on the solution of the neutron diffusion problem by an implicit [illegible] method . . . 314
R. Bellman: Functional equations and maximum range . . . 317
A. [illegible]: On the determination of certain thermodynamical physical quantities . . . 319
A. H. Zemanian: Correction to my paper "On transfer functions and transients" . . . 320

Book Reviews . . . 230, 262, 270, 284, 296, 321